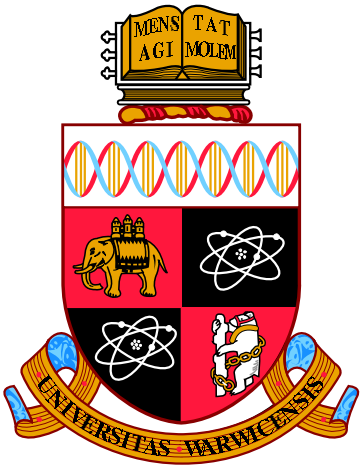


Finite Mixture Modeling with Non-Local Priors

by

Jairo Alberto Fúquene Patiño

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

September 2018

THE UNIVERSITY OF WARWICK

Contents

List of Tables **iv**

List of Figures **v**

Acknowledgments **ix**

Declarations **x**

Abstract **xi**

Abbreviations **xiv**

Chapter 1 Introduction **1**

1.1 Finite mixtures and properties 1
1.2 Parameter estimation in finite mixtures 4
1.2.1 Maximum likelihood estimation 5
1.2.2 Gibbs sampling 5
1.3 Model selection strategies in mixtures 7
1.4 Non-local priors in the context of mixtures 10
1.5 Contributions in this thesis 15
1.6 Outline 16

Chapter 2 Theoretical framework **18**

2.1 A general NLP class for mixture distributions 19
2.1.1 Application to Normal and T mixtures 20
2.1.2 Application to Binomial mixtures 23
2.1.3 Application to product Binomial mixtures 23
2.2 Parsimony enforcement 24
2.2.1 Technical conditions 24
2.2.2 Additional conditions from Rousseau and Mengersen (2011) . 25

2.2.3 Discussion of the technical conditions 27
2.3 Theoretical characterization of the sparsity 28

Chapter 3 Prior computation and elicitation **30**
3.1 Prior normalization constant 30
3.2 Prior elicitation 32

Chapter 4 Computational framework **38**
4.1 Integrated likelihood 39
4.2 Posterior mode estimation 42
4.2.1 Derivation of the EM algorithm for Normal mixtures under MOM-IW-Dir priors 45
4.2.2 Derivation of the EM algorithm for product Binomial mixtures under MOM-Beta-Dir priors 47
4.3 Precision of MCMC-based estimates relative to local priors 49

Chapter 5 Simulation studies **54**
5.1 Normal mixture 54
5.2 Misspecified mixtures 61
5.3 Binomial mixture 68
5.4 Sensitivity to the prior 68
5.4.1 Sensitivity to choosing q 70
5.4.2 Sensitivity to choosing g for MOM-IW priors 70
5.4.3 Sensitivity to choosing g for MOM-Beta priors 70
5.5 An illustration of computations under product of Binomial mixtures 78

Chapter 6 Computationally-fast alternatives **81**
6.1 Exploration of non-local model selection criteria 81
6.2 Bayes factors for mixtures from cluster occupancies 84
6.2.1 Comparison with other alternatives 86
6.2.2 Computational cost and precision across MCMC runs for the ECP estimator 89

Chapter 7 Applications **96**
7.1 Old Faithful 96
7.2 Cytometry data 99
7.3 Fisher's Iris data 102
7.4 Comparison with overfitted and repulsive overfitted mixtures 104
7.5 Political blog data 105

Chapter 8 Conclusions and future work **110**

Appendix A Proofs **113**

A.1 Auxiliary lemmas to prove Theorem 1 113
A.2 Proof of Theorem 1 115
A.3 Proof of Lemma 1 118
A.4 Proof of Corollary 1 119
A.5 Proof of Corollary 2 121
A.6 Proof of Corollary 3 121
A.7 Proof of Proposition 2 122

Appendix B MCMC results **124**

Appendix C Probability density functions **138**

List of Tables

Table 3.1 Estimation of $\log(C_k)$ for the MOM-IW prior. 33
Table 3.2 Estimation of $\log(C_k)$ for the MOM-Beta prior. 34

Table 5.1 Cases for the simulation study data-generating truth. 55
Table 5.2 Misspecified mixtures. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ under Normal-IW-Dir, MOM-IW-Dir, BIC and sBIC. 64
Table 5.3 Product Binomial simulation. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ and $k^* = 4$ under Beta and MOM-Beta priors, BIC and AIC. 80

Table 6.1 Simulation study. Mean $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, 2, 3\}$ and Cases 1, 3, 5 and 7 under MOM-IW-Dir and Normal-IW-Dir priors. Median CPU time (seconds). 89

Table 7.1 Faithful dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ under Normal-IW-Dir, MOM-IW-Dir, BIC and BIC and sBIC. 99
Table 7.2 Cytometry dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ under Normal-IW-Dir, MOM-IW-Dir, BIC and BIC and sBIC. 100
Table 7.3 Iris dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ under Normal-IW-Dir, MOM-IW-Dir, BIC and BIC and sBIC. 102
Table 7.4 Posterior for the distribution on non-empty components in overfitted mixtures, the misspecified student-T mixture, Faithful, Iris and Cytometry data. 104
Table 7.5 Posterior distribution on non-empty components in repulsive overfitted mixtures. The misspecified student-T mixture, Faithful, Iris and Cytometry. 105
Table 7.6 Combined words for the USA political blog data. 107
Table 7.7 20 most representative words for each component. 107
Table 7.8 USA political blogs dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ under Beta-Dir, MOM-Beta-Dir, BIC and BIC. 109

List of Figures

Figure 1.1 The MOM prior density with $t = 1$ and for $g = \{1, 2, 3\}$. . . . 12

Figure 2.1 Default MOM-IW 22
Figure 2.2 Default MOM-Beta $p(\boldsymbol{\theta} \mid g = 7.05, \mathcal{M}_2)$ (left) and Beta $p^L(\boldsymbol{\theta} \mid g^L = 1.98, \mathcal{M}_2)$ (right) 23

Figure 3.1 Illustration of prior densities for η. 35
Figure 3.2 Illustration of prior densities for η with $p = \{1, 2, 3, 4\}$ for Normal mixtures with unequal covariances. 36

Figure 4.1 Precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ in 100 univariate simulations, $k^* = 1$. 50
Figure 4.2 Precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ in 100 bivariate simulations, $k^* = 1$. . 51
Figure 4.3 EM estimates for 300 data sets of $n = 500$ from a univariate Normal where $\mu = 0$ and $\sigma^2 = 1$. 52
Figure 4.4 EM estimates for 300 data sets of $n = 500$ from a bivariate three component Normal mixture where $\boldsymbol{\mu}_1 = (-1, -1)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (3, 3)'$, $\boldsymbol{\eta}$=(0.25,0.25,0.5), $\sigma^2_{11} = \sigma^2_{22} = 1$ and $\sigma^2_{12} = \sigma^2_{21} = -0.5$. 53

Figure 5.1 Simulation study data-generating truth. 56
Figure 5.2 Simulation study. Univariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n for the MOM-IW and Normal-IW. 57
Figure 5.3 Simulation study. Bivariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n for the MOM-IW and Normal-IW. 58
Figure 5.4 Simulation study. Univariate mixtures. Proportion of correct $\hat{k} = k$ vs. n for MOM-IW, Normal-IW, AIC and BIC. 59
Figure 5.5 Simulation study. Bivariate mixtures. Proportion of correct $\hat{k} = k$ vs. n for MOM-IW, Normal-IW, AIC and BIC. 60
Figure 5.6 Simulated data and contour in logarithm scale for the data-generating student-T mixture with $\upsilon_j = 4$. 61

Figure 5.7 Simulated data and contour in logarithm scale for the data-generating iskew-T mixture with $\upsilon_j = 100$. 62
Figure 5.8 Simulated data and contour in logarithm scale for the data-generating iskew-T mixture with $\upsilon_j = 4$. 63
Figure 5.9 Misspecified Student-T mixture. Estimated contours for (a) BIC/sBIC (top left), (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data. 65
Figure 5.10 Misspecified iskew-T mixture with $\upsilon_j = 100$. Estimated contours for (a) BIC (top left), (b) AIC/sBIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data. 66
Figure 5.11 Misspecified iskew-T mixture with $\upsilon_j = 4$. Estimated contours for (a) BIC/sBIC (top left), (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data. 67
Figure 5.12 Binomial mixture. Frequencies of $\hat{k}$ for BIC, $\overline{\text{sBIC}}$, $\overline{\text{sBIC}}_{05}$, Beta and MOM-Beta. 69
Figure 5.13 Simulation study. Univariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = p + 1$ for the MOM-IW-Dir and Normal-IW-Dir. 71
Figure 5.14 Simulation study. Univariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = 4$. 72
Figure 5.15 Simulation study. Bivariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = p + 1$ for the MOM-IW-Dir and Normal-IW-Dir. 73
Figure 5.16 Simulation study. Bivariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = 16.5$ for the MOM-IW-Dir and Normal-IW-Dir. 74
Figure 5.17 Simulation study. Univariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n under $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$ for the MOM-IW-Dir and Normal-IW-Dir. 75
Figure 5.18 Simulation study. Bivariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n under $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$ for the MOM-IW-Dir and Normal-IW-Dir. 76
Figure 5.19 Binomial mixture. Frequencies of $\hat{k}$ for MOM-Beta for $g = 7.05$, $g = 16.09$ and $g = 29.99$ with $q = 2$. 77

Figure 5.20 Product Binomial simulation. MCMC trace plots for θ_{jf} corresponding to components $j = 1, ..., 4$ and variables $f = 1, \ldots, 8$. The colours in the trace plots indicate the different components. 79

Figure 6.1 Normal mixture. Frequencies of $\hat{k}$ for Cases 1 to 4 for 100 data sets, sample sizes of $n = 250$, $n = 500$ and $n = 1000$ using (6.1.4) and MOM-IW priors. 83
Figure 6.2 Binomial mixture. Frequencies of $\hat{k}$ for 100 data sets. Left: the data considered in Section 5.3 with $k^* = 4$. Right: simulated data from $n = 50$, 200 and 500, $L_{if} = 30$, $\eta_j = 1/4$, $\theta_j = \{0.05, 0.35, 0.65, 0.95\}$ and $k^* = 4$. 84
Figure 6.3 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir using the Marin and Robert (2008) estimator and ECP estimator for $n = 200$ observations in simulation Case 1. 87
Figure 6.4 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir using the Marin and Robert (2008) estimator and ECP estimator for $n = 200$ observations in simulation Case 3. 88
Figure 6.5 Precision of $\hat{P}(\mathcal{M}_{k^*} \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. 90
Figure 6.6 Precision of $\hat{P}(\mathcal{M}_k^* \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. 91
Figure 6.7 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. 92
Figure 6.8 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. 93
Figure 6.9 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. 94
Figure 6.10 Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. 95

Figure 7.1 Old Faithful: the biggest cone-type geyser located in the Yellowstone National Park, Wyoming, United States. 96
Figure 7.2 Classification and contours for the model chosen by MOM-IW-Dir for Faithful data set. 97
Figure 7.3 Faithful dataset. Contours for the model chosen by (a) BIC and (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the data. 98
Figure 7.4 Cytometry data-set with the variables CD4 and CD8b. 100

Figure 7.5 Projection of the variables CD4 and CD8b for the Cytometry data-set, classification of observations and contours using EM algorithm for BIC and AIC (top), and under Normal-IW-Dir and MOM-IW-Dir (bottom) 101
Figure 7.6 Top: The species, setosa, versicolor and virginica in the Fisher's Iris data set. Bottom: Classification for the model chosen by MOM-IW-Dir for Iris data set. 103
Figure 7.7 Word cloud for the political blog data set where sizes are increasing with frequency of use. 106
Figure 7.8 Posterior cluster probabilities $p(z_i = j|\mathbf{y}, \mathcal{M}_j)$ under BIC, AIC and Beta-Dir, MOM-Beta-Dir for documents labelled as conservative or liberal 108
Figure 7.9 Political blog data. Each blog was assigned to its most probable component under a MOM-Beta prior. Word sizes based on chi-square residuals from cross-tabulating word frequency versus assigned component 109

Figure B.1 MCMC results for the faithful data set. 125
Figure B.2 MCMC results for the misspecified data set. 126
Figure B.3 MCMC results for the Iris data set. 127
Figure B.4 MCMC results for the Cytometry data set. 128
Figure B.5 MCMC results for the Cytometry data set. 129
Figure B.6 MCMC results for the Cytometry data set. 130
Figure B.7 MCMC results for the Cytometry data set. 131
Figure B.8 Additional MCMC results for the computations for the product of Binomial mixture under MOM-Beta priors. MCMC output for weight parameters. 132
Figure B.9 MCMC results for the Political blog data. MCMC output for the words, $\hat{\theta}_{jf}$. 133
Figure B.10 MCMC results for the Political blog data. MCMC output for the words, $\hat{\theta}_{jf}$. 134
Figure B.11 MCMC results for the Political blog data. MCMC output for the words, $\hat{\theta}_{jf}$. 135
Figure B.12 MCMC results for the Political blog data. MCMC output for the words, $\hat{\theta}_{jf}$. 136
Figure B.13 MCMC results for the Political blog data. MCMC output for the words $\hat{\theta}_{jf}$ and weight parameters. 137

Acknowledgments

I dedicate this thesis to God and my parents, Lilia Patiño and Alberto Fúquene.

I would like to thank my supervisors Mark F.J. Steel and David Rossell for their strong support during these four and a half years of advice and for giving me the opportunity to learn in each step with their professional experience.

I thank professors Isobel Claire Gormley and Jim Smith, my revision panel, for the constructive comments and invaluable feedback.

Many thanks to Julieth Castañeda, Brenda Betancourt, Mafe Fúquene, Sandra Fúquene, Jorge Fúquene and Giovanny Fúquene, for your unconditional support.

I acknowledge the department of Statistics at the University of Warwick who funded my PhD studies.

Declarations

This thesis is a result of my own work, which was performed between October of 2013 and September of 2018 under the supervision of Professors Mark F.J. Steel and David Rossell. The material in this thesis is original, except in those cases where indicated by references. This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented was carried out by the author except where explicitly indicated otherwise by references. Most materials of this thesis has formed the following paper:

On choosing mixture components via non-local priors. Jairo Fúquene, Mark Steel, David Rossell. Under revision for Journal of the Royal Statistical Society: Series B (Statistical Methodology). https://arxiv.org/pdf/1604.00314.pdf.

Abstract

Choosing the number of mixture components remains a central but elusive challenge. Traditional model selection criteria can be either overly liberal or conservative when enforcing parsimony. They may also result in poorly separated components of limited practical use. In this thesis, the term parsimony refers to selecting a simpler model by enforcing a separation between the models under consideration, and the term sparsity refers to the ability of penalizing overffited models leading to well-separated components with non-negligible weight, interpretable as distinct subpopulations. Non-local priors (NLPs) are a family of distributions that encourage parsimony by enforcing a separation between the models under consideration. In this thesis we investigate the use of NLPs to choose the number of components in mixture models. Our main contributions are proposing the use of non-local priors (NLPs) to select the number of components, characterizing the properties of the associated inference (in particular, improved sparsity) and proposing tractable expressions suitable for prior elicitation purposes, simpler and computationally efficient algorithms and practical applications.

Chapter 2 develops the theoretical framework. We present NLPs in the context of mixtures and show how they lead to well-separated components that have non-negligible weight, hence interpretable as distinct subpopulations. Moreover we formulate a general NLP class, propose a particular choice leading to tractable expressions and give a theoretical characterization of the sparsity induced by NLPs for choosing the number of mixture components. Although the framework is generic we fully develop multivariate Normal, Binomial and product Binomial mixtures

based on a family of exchangeable moment priors.

Chapter 3 presents the prior computation and elicitation. We suggest default prior settings based on detecting multi-modal Normal and T mixtures, and minimal informativeness for categorical outcomes where multi-modality is not a natural consideration. The theory and underlying principles in this thesis hold more generally as outlined in Chapter 2, however.

Chapter 4 presents the computational framework for model selection and fitting. We propose simple algorithms based on Markov chain Monte Carlo methods and Expectation Maximization algorithms to obtain the integrated likelihood and parameter estimates.

Chapters 5-7 contain the simulation studies and applications. In Chapter 5 we compare the performance of our proposal to its local prior counterpart as well as the Bayesian Information Criterion (BIC), the singular Bayesian Information Criterion (sBIC) and the Akaike Information Criterion (AIC). Our results show a serious lack of sensitivity of the Bayesian information criterion (BIC) and insufficient parsimony of the AIC and the local prior counterpart to our formulation. The singular BIC behaved like the BIC in some examples and the AIC in others.

In Chapter 6 we explore a computational fast non-local model selection criteria and propose a new computational strategy that provides a direct connection between cluster occupancies and Bayes factors with the advantage that Bayes factors allow for more general model comparisons (for instance equal vs unequal covariances in Normal mixtures). This new computational strategy is helpful to discard unoccupied clusters in overfitted mixtures and we remark that the result has interest beyond purely computational purposes, e.g. to set thresholds on empty cluster probabilities in overfitted mixtures.

In Chapter 7 we present the applications of this thesis and also offer comparisons to overfitted and repulsive overfitted mixtures. In most examples their performance was competitive but depended on setting the prior parameters adequately to prevent the appearance of spurious components. The number of components inferred under NLPs was closer to the true number (when this was known) and

remained robust to prior parameter changes, provided these remain in the range of recommended defaults.

In Chapter 8 we have the conclusions and some possible future directions of this work. Finally, in Appendix A we present the proofs of Theorem 1 as well as auxiliary lemmas and corollaries. Appendix B shows the MCMC diagnostics. Appendix C presents the main probability density functions used throughout this thesis.

Abbreviations

- LPs: Local priors
- NLPs: Non-local priors
- MOM-IW: Moment Inverse Wishart
- MOM-Beta: Moment Beta
- T: Student-t
- EM: Expectation Maximization
- MCMC: Markov chain Monte Carlo
- BIC: Bayesian Information Criterion
- sBIC: Singular Bayesian Information Criterion
- AIC: Akaike Information Criterion
- MLE: Maximum Likelihood Estimator

Chapter 1

Introduction

Mixture models have many applications, *e.g.* in human genetics (Schork et al., 1996), false discovery rate control (Efron, 2008), signal deconvolution (West and Turner, 1994), density estimation (Escobar and West, 1995) and cluster analysis (*e.g.* Fraley and Raftery (2002); Baudry et al. (2012)). An extensive treatment is provided in Frühwirth-Schnatter (2006) and Mengersen et al. (2011). In spite of their fundamental role in statistics, due to their irregular nature (*e.g.* multi-modal unbounded likelihood, non-identifiability) choosing the number of components remains an elusive problem both in the Bayesian and frequentist paradigms.

As discussed below, despite the fact that formal criteria may achieve model selection consistency as the sample size grows to infinity (Gassiat and Handel, 2013), in practice they may lead to too many or too few components and require the data analyst to perform some ad-hoc post-processing. In this chapter we present an overview of finite mixture distributions. Section 1.1 presents the finite mixture distributions and their properties. Section 1.2 contains some alternatives for parameter estimation. In Sections 1.3 and 1.4 we discuss model selection strategies for mixtures and present NLPs in the context of mixture distributions, respectively.

1.1 Finite mixtures and properties

We consider a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ of independent observations from a finite mixture where $\mathbf{y}_i \in \mathbb{R}^p$ arises from the density

$$p(\mathbf{y}_i \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \sum_{j=1}^{k} \eta_j p(\mathbf{y}_i \mid \boldsymbol{\theta}_j). \tag{1.1.1}$$

The component densities $p(\mathbf{y}_i \mid \boldsymbol{\theta}_j)$ are indexed by a parameter $\boldsymbol{\theta}_j \in \Theta$, $\boldsymbol{\eta} = (\eta_1, ..., \eta_k) \in \mathcal{E}_k$ denotes the weights, $\mathcal{E}_k$ the unit simplex and $\mathcal{M}_k$ the model with k components.

In this work, our main goal is to infer k. For simplicity we assume that there is an upper bound K such that $k \in \{1, \ldots, K\}$, *e.g.* given by subject-matter or practical considerations, but our framework remains valid by setting a prior distribution on k with support on the natural numbers.

We denote the whole parameter set as $\boldsymbol{\vartheta}_k = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Theta_k = \Theta^k \times \mathcal{E}_k$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$. We assume that the sample $\mathbf{y}$ is truly generated by $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$ for some $k^* \in \{1, \ldots, K\}$, $\boldsymbol{\vartheta}^*_{k^*} \in \Theta_{k^*}$. Model (1.1.1) can be equivalently formulated in terms of latent cluster allocations $\mathbf{z}$ given by

$$z_{ij} = \begin{cases} 1 & \text{if } i \text{ belongs to component } j, \\ 0 & \text{otherwise,} \end{cases} \tag{1.1.2}$$

and the complete-data likelihood given by

$$p(\mathbf{y}_i \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \prod_{i=1}^{n} \prod_{j=1}^{k} (\eta_j p(\mathbf{y}_i \mid \boldsymbol{\theta}_j))^{z_{ij}}. \tag{1.1.3}$$

As a first example, we consider a mixture of Normal distributions where the component densities (the main probability density functions used throughout this thesis are presented in Appendix C) are

$$p(\mathbf{y} \mid \boldsymbol{\theta}_j) = \mathrm{N}(\mathbf{y}; \boldsymbol{\mu}_j, \Sigma_j) \tag{1.1.4}$$

with $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$ where $\boldsymbol{\mu}_j \in \mathbb{R}^p$ is the mean and Σ_j the covariance matrix of component j. As a second example, we consider mixtures of heavy-tailed alternatives such as Student-t densities

$$p(\mathbf{y} \mid \boldsymbol{\theta}_j) = \mathrm{T}(\mathbf{y}; \boldsymbol{\mu}_j, \Sigma_j, \upsilon_j), \tag{1.1.5}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j, \upsilon_j)$ and υ_j are the degrees of freedom.

Another class of mixture distributions we use in this work is the product Binomial mixtures with mass function

$$p(\mathbf{y}_i \mid \boldsymbol{\theta}_j) = \prod_{f=1}^{p} \binom{L_{if}}{y_{if}} \theta_{jf}^{y_{if}} (1 - \theta_{jf})^{L_{if} - y_{if}}, \tag{1.1.6}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ are the number of successes observed for individual i across p outcomes, L_{if} the number of trials, θ_{jf} is the success probability for outcome f under component j, and $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jp})$. In particular the case $p = 1$ corresponds to a Binomial mixture.

As Frühwirth-Schnatter (2006) pointed out, one potential cause for the lack of identifiability is caused by the invariance of the likelihood (1.1.1) to relabeling the components. In the case of a mixture distribution with k components we have $k!$ equivalent ways of arranging the components. Consider for example the following subset $\mathcal{J}^P(\boldsymbol{\vartheta}) \subset \Theta_k$:

$$\mathcal{J}^P(\boldsymbol{\vartheta}) = \bigcup_{\psi \,\in\, \mathfrak{N}(k)} \{\boldsymbol{\vartheta}^* \,\in\, \Theta_k : \boldsymbol{\vartheta}^* = \psi(\boldsymbol{\vartheta})\}, \tag{1.1.7}$$

where $\mathfrak{N}(k)$ denotes the set of the $k!$ permutations of $\{1, .., k\}$ and ψ is one of those permutations. In (1.1.7), $\boldsymbol{\vartheta}$ and any point $\boldsymbol{\vartheta^*} \in \mathcal{J}^P(\boldsymbol{\vartheta})$ generates the same distribution for $\mathbf{y}_i$. Relabeling (also known as label switching) is due to there being $k!$ equivalent ways of rearranging the components giving rise to the same $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$. Although it creates some technical difficulties, it does not seriously hamper inference. For instance, if $k = k^*$ then the maximum likelihood estimator (MLE) is consistent and asymptotically Normal as $n \to \infty$ in the quotient topology (Redner, 1981), and from a Bayesian perspective the integrated likelihood behaves asymptotically as in regular models (Crawford, 1994).

To illustrate an over-fitted mixture (Frühwirth-Schnatter, 2006), consider a three-component mixture as follows

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\vartheta}_2, \mathcal{M}_3) &= \eta_1 p(\mathbf{y}|\boldsymbol{\theta}_1) + \eta_2 p(\mathbf{y}|\boldsymbol{\theta}_2) + 0 p(\mathbf{y}|\boldsymbol{\theta}_3) && (1.1.8) \\ &= \eta_1 p(\mathbf{y}|\boldsymbol{\theta}_1) + (\eta_2 - \eta_3) p(\mathbf{y}|\boldsymbol{\theta}_2) + \eta_3 p(\mathbf{y}|\boldsymbol{\theta}_2). && (1.1.9) \end{aligned}$$

The set $S_0 = \{\boldsymbol{\vartheta}_3 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*, \eta_1 = \eta_1^*, \eta_2 = \eta_2^*, \eta_3 = 0\}$ is a non-identifiability set, the density $p(\mathbf{y}|\boldsymbol{\vartheta}_2, \mathcal{M}_3)$ is the same for arbitrary values $\boldsymbol{\theta}_3$ in S_0. The same situation is presented for the set $S_1 = \{\boldsymbol{\vartheta}_3 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*, , \boldsymbol{\theta}_3 = \boldsymbol{\theta}_2^*, \eta_1 = \eta_1^*\}$, as the density $p(\mathbf{y}|\boldsymbol{\vartheta}_2, \mathcal{M}_3)$ is the same for arbitrary values η_3 with $0 \leq \eta_3 \leq \eta_2$. Non-identifiability due to overfitting has more serious consequences, *e.g.* estimates for $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ are consistent under mild conditions (Ghosal and der Vaart, 2001) but the MLE and posterior mode of $\boldsymbol{\vartheta}_k$ can behave erratically (Leroux, 1992; Rousseau and Mengersen, 2011; Ho and Nguyen, 2016). In addition, as we now discuss, frequentist and Bayesian tests to assess the adequacy of $\mathcal{M}_k$ can behave unsatisfactorily.

1.2 Parameter estimation in finite mixtures

Parameter estimation in mixture distributions is challenging because of the lack of identifiability. There is an extensive literature on computational methods to estimate parameters in mixture distributions (see for example Neal (1996), Crawford (1994), Frühwirth-Schnatter (2011), Frühwirth-Schnatter (2006) and Mengersen et al. (2011)). Chapters 4 and 6 we present two algorithms to estimate the integrated likelihood from MCMC output. In Section 4.1 of Chapter 4 we show the first one proposed by Marin and Robert (2008) and, while we found it to be reasonably accurate, it is limited to conjugate models and requires an MCMC post-processing step that may have non-negligible cost. In Section 6.2 of Chapter 6 the second algorithm is novel (to our knowledge), applicable to non-conjugate models and only requires cluster probabilities available as an MCMC by-product, avoiding costly post-processing. The algorithm proposed by Marin and Robert (2008) uses posterior samples obtained from a Gibbs Sampling algorithm referred to as data augmentation (Frühwirth-Schnatter (2006)). This algorithm uses iterative steps for sampling the parameters of the mixture from their full conditional distributions by defining a missing data structure of the data. Although the Gibbs sampling algorithm is feasible, practical, and applicable to many mixture distributions for posterior inference purposes, EM algorithms (Dempster et al. (1977)) could also be a fast alternative to obtain posterior modes. For a Gibbs sampling algorithm we iteratively take samples of the parameters $\mathbf{z}^{(t)}$, $\boldsymbol{\theta}_j^{(t)}$ and $\eta_j^{(t)}$ from the full conditional distributions, $p(\mathbf{z}^{(t)}|\boldsymbol{\theta}_j^{(t-1)}, \eta_j^{(t-1)}, \mathcal{M}_k)$, $p(\boldsymbol{\theta}_j^{(t)}|\mathbf{z}^{(t)}, \eta_j^{(t-1)}, \mathcal{M}_k)$ and $p(\eta_j^{(t)}|\boldsymbol{\theta}_j^{(t)}, \mathbf{z}^{(t)}, \mathcal{M}_k)$, respectively. In the EM algorithm, for the E-step and the t-th iteration we compute the conditional expectations of the missing data variables, $z_{ij}^{(t)}$ for $i = 1, ..., n$, given the data and the current parameters $\boldsymbol{\theta}_j^{(t-1)}$ and $\eta_j^{(t-1)}$. In the M-step and the t-th iteration we compute the maximizers $\boldsymbol{\theta}_j^{(t)}$ and $\eta_j^{(t)}$ of the logarithm of (1.1.3) given the expectations of the missing data.

As an illustration, in Sections 1.2.1 and 1.2.2 we outline the Gibbs sampling and EM algorithms (Dempster et al. (1977)) for Normal and product Binomial mixtures given in (1.1.4).

1.2.1 Maximum likelihood estimation

EM for Normal mixtures

For $t \geq 1$ and $j = 1, ..., k$ given $\boldsymbol{\vartheta}_j^{(0)} = (\boldsymbol{\mu}_1^{(0)}, ...\boldsymbol{\mu}_k^{(0)}, \Sigma_1^{(0)}, ..., \Sigma_k^{(0)}, \boldsymbol{\eta}^{(0)})$ in the E-step we use

$$\bar{z}_{ij}^{(t)} = p(z_{ij} = 1|\mathbf{y}_i, \boldsymbol{\vartheta}_j^{(t-1)}) = \frac{\eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\mu}_{\boldsymbol{k}}^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^k \eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\mu}_{\boldsymbol{j}}^{(t-1)}, \Sigma_j^{(t-1)})}.$$

In the M-step the estimators of the component means are

$$\boldsymbol{\mu}_j^{(t)} = \frac{\sum_{i=1}^n \bar{z}_{ij}^{(t)} \mathbf{y}_i}{\sum_{i=1}^n \bar{z}_{ij}^{(t)}},$$

and the component weights and variance-covariance matrix estimates are respectively given by

$$\eta_j^{(t)} = \frac{\sum_{i=1}^n \bar{z}_{ij}^{(t)}}{n}; \qquad \Sigma_j^{(t)} = \frac{\sum_{i=1}^n \bar{z}_{ij}^{(t)} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})'}{\sum_{i=1}^n \bar{z}_{ij}^{(t)}}.$$

In the case of a common variance-covariance matrix $\Sigma_j = \Sigma$, the estimates

$$\Sigma^{(t)} = \frac{\sum_{j=1}^k \sum_{i=1}^n z_{ij}^{(t)} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})'}{n}.$$

EM for product Binomial mixtures

For $t \geq 1$ and $j = 1, ..., k$ given $\boldsymbol{\vartheta}_j^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ...\boldsymbol{\theta}_k^{(0)}, \boldsymbol{\eta}^{(0)})$ in the E-step we use

$$\bar{z}_{ij}^{(t)} = \frac{\eta_j^{(t-1)} \prod_{f=1}^p \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}{\sum_{j=1}^k \eta_j^{(t-1)} \prod_{f=1}^p \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}.$$

For the M-step the estimators of the component probabilities and the component weights are respectively as follows

$$\theta_{jf} = \frac{\sum_{i=1}^n \bar{z}_{ij}^{(t)} y_{if}}{\sum_{i=1}^n \bar{z}_{ij}^{(t)} (L_{if} - y_{if})}, \qquad \eta_j^{(t)} = \frac{\sum_{i=1}^n \bar{z}_{ij}^{(t)}}{n}.$$

1.2.2 Gibbs sampling

Gibbs sampling for Normal mixtures

We start with some initial values for the parameters $\boldsymbol{\vartheta}_j^{(0)} = (\boldsymbol{\mu}_1^{(0)}, ..., \boldsymbol{\mu}_k^{(0)}, \Sigma_1^{(0)}, ..., \Sigma_k^{(0)}, \boldsymbol{\eta}^{(0)})$ and repeat the following steps for $t = 1, .., T$. Consider a prior that factors across

parameter components $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \prod_{j=1}^{k} N(\boldsymbol{\mu}_j \mid \mathbf{0}, g\Sigma_j)\,\mathrm{IW}(\Sigma_j \mid \nu, S)\mathrm{Dir}(\boldsymbol{\eta} \mid q)$, where g is a known scale, then iteratively

S1 Sample $z_{ij}^{(t)}$ from its conditional posterior distribution as follows, for $j = 1, ..., k$

$$p(z_{ij}^{(t)} = 1|\boldsymbol{\vartheta}_j^{(t-1)}, \mathbf{y}_i) = \frac{\eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} p(\mathbf{y}_i|\boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})},$$

S2 Conditional on the classification $\boldsymbol{z}^{(t)}$,

a1 Sample $\boldsymbol{\eta}^{(t)}|\boldsymbol{\mu}_1^{(t-1)}, ..., \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{z}^{(t)}, \Sigma_1^{(t-1)}, ..., \Sigma_k^{(t-1)}, \mathbf{y}_1, ..., \mathbf{y}_n$ using a Dirichlet distribution

$$\boldsymbol{\eta}^{(t)} \sim \mathrm{Dir}(q + n_1^{(t)}, ..., q + n_k^{(t)}),$$

where $q + n_1^{(t)}, ..., q + n_k^{(t)}$ are the hyperparameters of a Dirichlet distribution and $n_j^{(t)} = \sum_{i=1}^{n} \mathrm{I}(z_{ij}^{(t)} = 1)$ is the number of observations assigned to the j-*th* component.

a2 Sample the variance-covariance matrix, $\Sigma_j^{(t)}|\boldsymbol{\mu}_1^{(t-1)}, ..., \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{z}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{y}_1, ..., \mathbf{y}_n$ from an Inverse Wishart distribution, for $j = 1, ..., k$, such that

$$\Sigma_j^{(t)} \sim \mathrm{IW}\left(\nu + n, S_j\right),$$

with $S_j = S^{-1} + \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})' + \dfrac{n_j/g}{n_j + 1/g} \bar{\mathbf{y}}_j \bar{\mathbf{y}}_j'$ where $\bar{\mathbf{y}}_j = \dfrac{1}{n_j} \sum_{i=1}^{n} z_{ij}^{(t)} \mathbf{y}_i$ and g is a known scale. For a common variance-covariance matrix $\Sigma_j = \Sigma$ we use $S_j = S^{-1} + \sum_{j=1}^{k} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t-1)})' + \sum_{j=1}^{k} \dfrac{n_j/g}{n_j + 1/g} \bar{\mathbf{y}}_j \bar{\mathbf{y}}_j'$.

a3 Sample $\boldsymbol{\mu}_j^{(t)}|\boldsymbol{\eta}^{(t)}, \boldsymbol{z}^{(t)}, \Sigma_1^{(t-1)}, ..., \Sigma_k^{(t-1)}, \mathbf{y}_1, ..., \mathbf{y}_n$ from a multivariate Normal distributions as follows, for $j = 1, ..., k$

$$\boldsymbol{\mu}_j^{(t)} \sim N\left(\frac{g(\sum_{i=1}^{n} z_{ij}^{(t)} \mathbf{y}_i)}{1 + g n_j^{(t)}}, \frac{g}{1 + g n_j^{(t)}} \Sigma_j^{(t)}\right).$$

Finally, some draws are discarded using a burn-in period.

Gibbs sampling for product Binomial mixtures

We use initial values for the parameters $\boldsymbol{\vartheta}_j^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)}, \boldsymbol{\eta}^{(0)})$ and repeat the following steps for $t = 1, .., T$. Consider a prior that factors across parameter com-

ponents $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \prod_{j=1}^{k}\prod_{f=1}^{p} \text{Beta}(\theta_{jf}; ag, (1-a)g)\text{Dir}(\boldsymbol{\eta} \mid q)$ then iteratively

S1 Sample $z_{ij}^{(t)}$ from its conditional posterior distribution as follows

$$p(z_{ij}^{(t)} = 1|\boldsymbol{\vartheta}_j^{(t-1)}, \mathbf{y}_i) = \frac{\eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}.$$

S2 Conditional on the classification $\boldsymbol{z}^{(t)}$,

a1 Sample $\boldsymbol{\eta}^{(t)}|\boldsymbol{\theta}_1^{(t-1)}, ..., \boldsymbol{\theta}_k^{(t-1)}, \boldsymbol{z}^{(t)}, \mathbf{y}_1, ..., \mathbf{y}_n$ using a Dirichlet distribution

$$\boldsymbol{\eta}^{(t)} \sim \text{Dir}(q + n_1^{(t)}, ..., q + n_k^{(t)}).$$

where $n_j^{(t)} = \sum_{i=1}^{n} \text{I}(z_i^{(t)} = j)$.

a2 Sample $\theta_{jf}^{(t)}|\boldsymbol{\eta}^{(t)}, \boldsymbol{z}^{(t)}, \mathbf{y}_1, ..., \mathbf{y}_n$ using a Beta distribution

$$\theta_{jf}^{(t)} \sim \text{Beta}\left(ag + \sum_{z_i^{(t)}=j} y_{if}, (1-a)g + \sum_{z_i^{(t)}=j} (L_{if} - y_{if})\right).$$

Finally, some draws are discarded using a burn-in period.

1.3 Model selection strategies in mixtures

The literature on criteria to choose k is extensive (see for example Richardson and Green (1997), Fraley and Raftery (2002), Baudry et al. (2012) and Gassiat and Handel (2013)). From a frequentist perspective the likelihood ratio test between $\mathcal{M}_k$ and $\mathcal{M}_{k+1}$ may diverge as $n \to \infty$ when data truly arise from $\mathcal{M}_k$ unless restrictions on the parameters or likelihood penalties are imposed (Ghosh and Sen (1985); Liu and Shao (2004); Chen and Li (2009)).

As an alternative one may consider criteria such as the Bayesian information criterion (BIC), Akaike's information criterion (AIC), the integrated complete likelihood (Biernacki et al., 2000) or the singular BIC (Drton and Plummer (2017), sBIC). The formal BIC justification as an approximation to the Bayesian evidence (Schwarz, 1978) is not valid for overfitted mixtures, however it is often adopted as a useful criterion (Fraley and Raftery, 2002). Other alternatives employed in Bayesian settings are the deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) and implemented in finite mixtures by Celeux et al. (2006), and the

BIC-MCMC (Mengersen et al. (2011), Chapter 10) obtained from the largest log-likelihood mixture across the MCMC draws. However, we study the performance of our approach with respect to BIC, AIC and sBIC, as these are more closely related to our proposal.

One issue is that the BIC ignores that $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ has $k!$ maxima, causing a loss of sensitivity to detect truly present components. More importantly, the dimensionality penalty $p_k = \dim(\Theta_k)$ used by the BIC is too large for overfitted mixtures (Watanabe, 2013), again decreasing power. These theoretical observations align with the empirical results we present here. The sBIC builds on Watanabe (2009, 2013) to improve the asymptotic approximation of the integrated likelihood. In our results the sBIC over-penalized model complexity in some examples (albeit less so than the BIC) but under-penalized in others, where it gave similar results to the AIC.

From a Bayesian perspective, model selection is usually based on the posterior probability $P(\mathcal{M}_k \mid \mathbf{y}) = p(\mathbf{y} \mid \mathcal{M}_k)P(\mathcal{M}_k)/p(\mathbf{y})$, where $P(\mathcal{M}_k)$ is the prior probability model,

$$p(\mathbf{y} \mid \mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) d\boldsymbol{\vartheta}_k \tag{1.3.1}$$

the integrated (or marginal) likelihood and $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ a prior distribution under $\mathcal{M}_k$. One may also use Bayes factors $B_{k',k}(\mathbf{y}) = p(\mathbf{y} \mid \mathcal{M}_{k'})/p(\mathbf{y} \mid \mathcal{M}_k)$ to compare any pair $\mathcal{M}_{k'}, \mathcal{M}_k$. A common argument for (1.3.1) is that it automatically penalizes overly complex models, however this parsimony is not as strong as one would ideally wish. To gain intuition, for regular models with fixed p_k one obtains

$$\log p(\mathbf{y} \mid \mathcal{M}_k) = \log p(\mathbf{y} \mid \hat{\boldsymbol{\vartheta}}_k, \mathcal{M}_k) - \frac{p_k}{2} \log(O_p(n)) + O_p(1) \tag{1.3.2}$$

as $n \to \infty$ (Dawid, 1999). This implies that $B_{k^*,k}(\mathbf{y})$ grows exponentially as $n \to \infty$ when $\mathcal{M}_{k^*} \not\subset \mathcal{M}_k$ but is only $O_p(n^{-(p_k - p_{k^*})/2})$ when $\mathcal{M}_{k^*} \subset \mathcal{M}_k$. That is, overfitted models are only penalized at a slow polynomial rate. Key to the current manuscript, Johnson and Rossell (2010) showed that either faster polynomial or quasi-exponential rates are obtained by letting $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ be a NLP (defined below). Expression (1.3.2) remains valid for many mixtures with $k \leq k^*$ (*e.g.* including Normal mixtures, Crawford (1994)), however this is no longer the case for $k > k^*$. Using algebraic statistics, Watanabe (2009, 2013) gave expressions analogous to (1.3.2) for overfitted $k > k^*$ where $p_k/2$ is replaced by a rational number $\lambda \in [p_{k^*}/2, p_k/2]$ called the *real canonical threshold* and the remainder term is

$O_p(\log\log n)$ instead of $O_p(1)$. The exact value of λ is complicated but the implication is that p_k in (1.3.2) imposes an overly stringent penalty that can decrease the sensitivity of the BIC, and also that the Bayes factor to penalize overfitted $k > k^*$ mixtures is $B_{k,k^*}(\mathbf{y}) = O_p(n^{-(\lambda - p_{k^*}/2)})$. That is, akin to regular models $k > k^*$ is penalized only at a slow polynomial rate. These results align with those in Chambaz and Rousseau (2008). Denoting the posterior mode by $\hat{k} = \arg\max_k P(\mathcal{M}_k \mid \mathbf{y})$, these authors found that the frequentist probability $P_{\vartheta^*_{k^*}}(\hat{k} < k^*) = O(e^{-an})$ but in contrast $P_{\vartheta^*_{k^*}}(\hat{k} > k^*) = O((\log n)^b/\sqrt{n})$ for some constants $a, b > 0$, again implying that overfitted mixtures are not sufficiently penalized.

We emphasize that these results apply to a wide class of priors but not to the NLP class proposed in this paper, for which faster rates are attained. Note also that the BIC and related likelihood penalties (where $\log(n)$ is replaced by a rate strictly between $\log\log(n)$ and n) attain consistency as $n \to \infty$ for fairly general mixtures (Gassiat and Handel, 2013), but as illustrated here for finite (potentially quite large) n the BIC can lack sensitivity.

An interesting alternative to considering $k \in \{1, \ldots, K\}$ is to set a single large k and induce posterior shrinkage, a strategy often referred to as *overfitted mixtures*. Rousseau and Mengersen (2011) showed that the prior on the weights $p(\boldsymbol{\eta} \mid \mathcal{M}_k)$ strongly influences posterior inference when $k > k^*$. Under $p(\boldsymbol{\eta} \mid \mathcal{M}_k) = \text{Dir}(\boldsymbol{\eta}; q_1, ..., q_k)$ with $\max_j q_j < d/2$ where $d = \dim(\Theta)$ the posterior of $\boldsymbol{\eta}$ collapses to 0 for redundant components, but if $\min_j q_j > d/2$ then it collapses on a solution where at least two components $i \neq j$ have identical parameters $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ and non-zero weights $\eta_i > 0$, $\eta_j > 0$. That is, the posterior shrinkage induced by $q_j < d/2$ helps discard spurious components. Gelman et al. (2013) set $q_1 = ... = q_k = 1/k$, but Havre et al. (2015) argued that this leads to insufficient shrinkage and proposed smaller q_j. One may then count the number of empty components at each Markov Chain Monte Carlo (MCMC) iteration to estimate k^*.

Petralia et al. (2012) argued that faster shrinkage may be obtained via overfitted repulsive priors, i.e. assigning vanishing density to $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ for $i \neq j$. Affandi et al. (2013) and Xu et al. (2016) gave related determinantal point process frameworks, and Xie and Xu (2017) proposed extensions to non-parametric Gaussian mixtures. A recent approach by Malsiner-Walli et al. (2017) resembling repulsive mixtures is to encourage nearby components merging into groups at a first hierarchical level and to then enforce between-group separation at the second level.

In spite of their usefulness, overfitted mixtures (whether repulsive or not) also bear limitations. Therefore, selecting the number of components in a mixture using the NLP approach compared to the overfitted or repulsive overfitted mixture

approach has several advantages:

(i) On the practical side one can study the number of components but cannot address more general model selection questions, say choosing equal versus different component-specific covariances. Also, inference may be sensitive to the chosen values of q_j, k, or the threshold to discard unoccupied components (see Chapter 7).

(ii) In terms of interpretation, cluster occupancy probabilities given by overfitted mixtures are different from model probabilities $p(\mathcal{M}_k \mid \mathbf{y})$ under NLPs. We compute posterior probabilities under a uniform model prior (i.e., equal prior model probabilities) having into account the uncertainty under all considered models. We remark that even though estimating $p(\mathcal{M}_k \mid \mathbf{y})$ requires one to consider multiple k, relative to overfitted mixtures where one sets a single large k, this can be handled as an embarrassingly parallel problem.

(iii) From a methodological view point, in Chapter 6 we show that Bayes factors, and hence $p(\mathcal{M}_k \mid \mathbf{y})$, are given by ratios of posterior to prior empty cluster probabilities. The result motivates a novel empty-cluster probability (ECP) estimator to obtain $p(\mathcal{M}_k \mid \mathbf{y})$ from standard MCMC output that is computationally-convenient and applicable to very general mixtures, both under LPs and NLPs.

The latter observation is conditional on adopting a careful prior elicitation, which is an important contribution of this thesis (see Chapter 3). We show that obtaining $p(\mathbf{y} \mid \mathcal{M}_k)$ under NLP is no harder than for local priors and easy to implement given MCMC output from standard local priors (see Chapter 4).

1.4 Non-local priors in the context of mixtures

To illustrate NLPs, let δ be the parameter of interest in a generic test for two hypothesis,

$$H_0 : \delta \in \Delta_0; \tag{1.4.1}$$

$$H_1 : \delta \in \Delta_1, \tag{1.4.2}$$

where $\Delta_0 \cup \Delta_1 = \Delta$. From a Bayesian perspective, we need to specify prior distributions $p(\delta|H_0)$ and $p(\delta|H_1)$ on δ under each hypothesis H_0 and H_1, respectively. Usually, Bayesian hypothesis tests are defined with LPs and $p(\delta|H_1)$ is a continuous density and positive on Δ_0. The main criticism in Johnson and Rossell (2010)

is that LPs do not incorporate a minimal notion of separation between both null and alternative hypothesis. They called a density a local prior if it is a continuous density satisfying

$$p(\delta|H_1) > \epsilon \quad \text{for all} \quad \delta \in \Delta_0. \tag{1.4.3}$$

On the other hand, if for every $\epsilon > 0$ there is a $\zeta > 0$ such that

$$p(\delta|H_1) < \epsilon \quad \text{for all} \quad \delta \in \Delta : \inf_{\delta \in \Delta_0} |\delta - \delta_0| < \zeta, \tag{1.4.4}$$

then $p(\delta|H_1)$ is called a NLP. Now we focus on a specific NLP class called moment (MOM) priors proposed in Johnson and Rossell (2010). Let $p_{\mathrm{b}}(\delta)$ be a base prior density with $2t$ finite integer moments ($t \geq 1$) and where $\delta \in \mathbb{R}$. The t-th MOM prior density, for a point null hypothesis $H_0 : \delta = \delta_0$, is defined as follows:

$$p_{\mathrm{M}}(\delta|H_1) = \frac{(\delta - \delta_0)^{2t}}{g\tau_t} p_b(\delta), \tag{1.4.5}$$

where

$$\tau_t = \int_{\Delta} \frac{(\delta - \delta_0)^{2t}}{g} p_{\mathrm{b}}(\delta) d\delta, \tag{1.4.6}$$

with δ_0 is a fixed value and g a known scale.

Figure 1.1 illustrates the MOM prior density with $t = 1$ and for $g = \{1, 2, 3\}$ with $p_{\mathrm{b}}(\delta)$=Normal$(\delta; 0, g)$. For the test of a null hypothesis $H_0 : \delta = \delta_0$ against the composite alternative $H_1 : \delta \neq \delta_0$, Johnson and Rossell (2010) showed that under certain regularity conditions by using MOM priors the convergence rate of Bayes factors in favor of the alternative hypothesis is $\mathcal{O}_p(n^{-t-\frac{1}{2}})$, for a true null hypothesis. In contrast with the convergence $\mathcal{O}_p(n^{-\frac{1}{2}})$ obtained by using LPs. For a true alternative hypothesis, the Bayes factor in favor of the null hypothesis decreases exponentially fast. Therefore, MOM priors ameliorate the imbalance in convergence rates for the case where $\delta \in \mathbb{R}$. Johnson and Rossell (2010) also studied multivariate extensions of MOM priors. They defined the multivariate MOM prior given by

$$p_M(\boldsymbol{\delta}) = \frac{Q(\boldsymbol{\delta})^t p_b(\boldsymbol{\delta})}{E_{p_b}[Q(\boldsymbol{\delta})^t]}, \tag{1.4.7}$$

with

$$Q(\boldsymbol{\delta}) = \frac{(\boldsymbol{\delta} - \boldsymbol{\delta}_0)' \Sigma^{-1} (\boldsymbol{\delta} - \boldsymbol{\delta}_0)}{n g_Q \sigma^2}, \tag{1.4.8}$$

where $\boldsymbol{\delta}$ is a $p \times 1$ dimensional real vector, Σ is a definite positive matrix, $g_Q > 0$

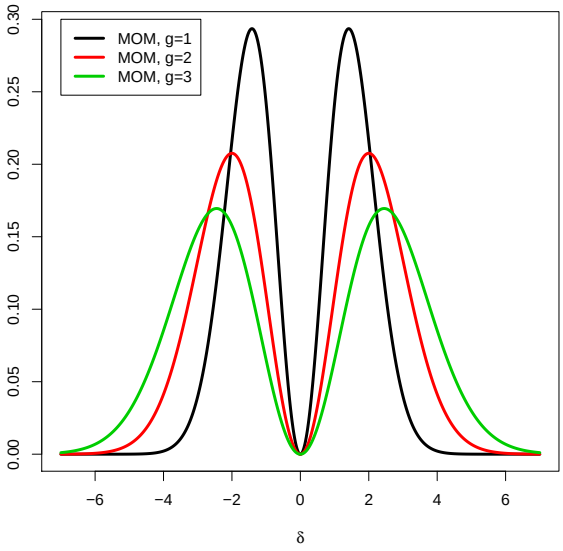


Figure 1.1: The MOM prior density with $t = 1$ and for $g = \{1, 2, 3\}$.

is a scalar, and $p_b(\boldsymbol{\delta}) > 0$ is a proper prior density on $\boldsymbol{\delta}$ with two bounded partial derivatives in a neighborhood containing $\boldsymbol{\delta}_0$ and $E_{p_b}[Q(\boldsymbol{\delta})^t]$ is finite. Johnson and Rossell (2010) showed that the multivariate MOM prior leads to a convergence of Bayes factor in favor of the alternative hypotheses equal to $\mathcal{O}_p(n^{-t-\frac{d}{2}})$ when the null hypothesis is true. In contrast with $\mathcal{O}_p(n^{-\frac{d}{2}})$ obtained by using LPs. In Johnson and Rossell (2010) NLPs are used for pairwise comparison of nested linear models, probit regression and test of variances.

Johnson and Rossell (2012) investigated the use of NLPs on the model regression parameters for comparing 2^p models, where p is the number of regressors. The authors showed that by using NLPs for a number of covariates $p = \mathcal{O}(n^\alpha)$, $\alpha < 1$ and $p < n$, the resulting model selection procedures are consistent in linear regression settings. Therefore, NLPs assign a posterior probability of 1 to the data generating model as the sample size $n \to \infty$ and under certain regularity conditions pertaining the design matrix. Johnson and Rossell (2012) showed that under the same conditions, Bayesian procedures with LPs lead to the asymptotic assignment of a posterior probability of 0, when $\alpha > 1/2$. They also found that their Bayesian procedure works as well or better than penalized likelihood methods. Let $\boldsymbol{y}_n = (y_1, ..., y_n)'$ be a random vector, $\boldsymbol{X}_n$ a $n \times p$ design matrix of real numbers, and $\boldsymbol{\beta}$ a $p \times 1$ regression vector with components β_i, $i = 1, \ldots, p$. Johnson and Rossell

(2012) explored models of the form

$$\boldsymbol{y}_n \sim N(\boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \tag{1.4.9}$$

with a MOM prior for $\boldsymbol{\beta}$ defined as follow

$$\begin{aligned} p(\boldsymbol{\beta}; \sigma^2, r, g) = d_p(2\pi)^{-p/2}(g\sigma^2)^{-rp-p/2}|\boldsymbol{A}_p|^{1/2} \\ \times\exp\{-\frac{1}{2g\sigma^2}\boldsymbol{\beta}'\boldsymbol{A}_p\boldsymbol{\beta}\}\prod_{i=1}^{p}\boldsymbol{\beta}_i^{2r}, \end{aligned} \tag{1.4.10}$$

for $g > 0$, $\boldsymbol{A}_p$ a $p \times p$ nonsingular scale matrix and $r \geq 1$ with r an integer. The normalizing constant d_p is independent of σ^2 and g. The NLP proposed in Johnson and Rossell (2012) given by 1.4.10 are different to the prior in equation (1.4.7) because they are product NLPs such as the priors we use in this thesis. As pointed out in Johnson and Rossell (2012), the multivariate MOM density presented in (1.4.7) is 0 only when all components of the parameter vector are 0. Therefore no penalty is induced on models that contain only a subset of 0 parameters. That is, (1.4.7) only separates the model with no variables from the full model. On the other hand, (1.4.10) is 0 if any component of the parameter vector is 0. Therefore the prior separates all 2^p models inducing a much stronger penalty on the regression parameters when any one of the vector components is 0.

Building upon Johnson and Rossell (2010, 2012), we formally define NLPs in the context of mixtures.

Definition 1 *Let $\mathcal{M}_k$ be the k-component mixture in (1.1.1). A continuous prior density $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ is a NLP iff*

$$\lim_{\boldsymbol{\vartheta}_k \to \mathbf{t}} p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = 0$$

for any $\mathbf{t} \in \Theta_k$ such that $p(\mathbf{y} \mid \mathbf{t}, \mathcal{M}_k) = p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k'}, \mathcal{M}_{k'})$ for some $\boldsymbol{\vartheta}_{k'} \in \Theta_{k'}$, $k' < k$.

A local prior (LP) is any $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ not satisfying Definition 1. Intuitively for nested $\mathcal{M}_{k'} \subset \mathcal{M}_k$ a NLP $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ penalizes any $\boldsymbol{\vartheta}_k$ that would be consistent with $\mathcal{M}_{k'}$, in our setting any k-mixture with redundant components. For instance an NLP under $\mathcal{M}_2$ must assign $p(\boldsymbol{\vartheta}_2 \mid \mathcal{M}_2) = 0$ whenever $p(\mathbf{y} \mid \boldsymbol{\vartheta}_2, \mathcal{M}_2)$ reduces to a one-component mixture, *e.g.* $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ or $\eta_1 \in \{0, 1\}$. That is one must penalize situations where two components have the same parameters (as in a repulsive mixture) and also when there are zero-weight components. This intuition is made precise in

Chapter 2 for the wide class of generically identifiable mixtures where $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ defines a NLP if and only if $\lim\ p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = 0$ as either:

(i) $\eta_j \to 0$ for any $j = 1, ..., k$.

(ii) $\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j$ for any $i \neq j$.

. Beyond their philosophical appeal in establishing a probabilistic separation between the models under consideration, Johnson and Rossell (2010) showed that for asymptotically Normal models NLPs penalize spurious parameters at a faster rate than (1.3.2), specifically depending on the speed at which $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ converges to 0. Johnson and Rossell (2012) found that NLPs are necessary and sufficient to achieve the strong consistency $P(\mathcal{M}_{k^*} \mid \mathbf{y}) \xrightarrow{P} 1$ in certain high-dimensional linear regression with $o(n)$ predictors, whereas Shin et al. (2018) showed a similar result with $o(e^n)$ predictors. These authors also observed gains in model selection relative to popular penalized likelihood methods.

1.5 Contributions in this thesis

Our main contribution in this work is proposing the use of non-local priors (NLPs) to select the number of mixture components. More specifically,

- We provide the *theoretical* characterization of the properties of the associated inference to choose the number of components (Chapter 2).

- We develop a *practical* framework placing emphasis on important aspects related to prior elicitation, and illustrate the framework in popular mixture families that include Normal, T, Binomial and product Binomial mixtures (Chapter 3).

- We build *computational* schemes proposing tractable expressions to compute the integrated likelihood and provide algorithms for posterior inference (Chapter 4).

- We study the performance of our proposal with respect to their local prior counterpart and BIC, sBIC, AIC, overfitted and repulsive overfitted mixtures using simulated and real data sets in Chapters 5 and 7.

- We also address the computational challenge of obtaining posterior model probabilities, both for LPs and NLPs. In Chapter 6 we propose an estimator based on showing that Bayes factors are ratios of posterior to prior empty-cluster probabilities. The estimator is applicable to a wide class of models and only requires empty-cluster probabilities, a natural by-product of MCMC algorithms. The result also helps set thresholds to drop unoccupied clusters in overfitted mixtures, and it is hence of independent interest.

1.6 Outline

This thesis is organized as follows. In Chapter 2 we present the theoretical aspects of this thesis. We formulate a general NLP class with a particular specification that leads to tractable expressions, and present applications to Normal, T and product Binomial mixtures. We discuss the technical conditions required to prove our main result given in Theorem 1 which states that the proposed NLP class leads to stronger parsimony than LPs.

In Chapter 3 we investigate the NLP in some detail and therefore develop methodology required for the proposed framework to be practical. We present how to compute the normalization constant for MOM priors avoiding a doubly intractable problem. Importantly, a natural elicitation for prior parameters is also proposed. This is a key issue in our setting as it defines what separation between components is deemed practically relevant for Normal and T mixtures, and addresses minimal informativeness for Binomial and product of Binomial mixtures.

In Chapter 4 we outline the computational schemes for model selection and parameter estimation. We present how to compute the integrated likelihood under MOM priors using an MCMC run from the posterior under local priors. We review some computational approximations of the integrated likelihood under LPs which are the current state-of-the-art. We also discuss posterior mode parameter estimates via an EM algorithm using a first order Taylor expansion of the penalty term. In the last section of this chapter we show comparisons of the proposed computational methods with existing approaches.

In Chapter 5 we illustrate the performance of our MOM-Inverse Wishart (MOM-IW) and MOM-Beta priors. We present a simulation study for univariate and bivariate Normal mixtures and compare the performance with respect to BIC and AIC. To illustrate the sBIC performance, a Binomial mixture example considered in Drton and Plummer (2017) is reproduced. We illustrate the use of MOM-IW in the presence of model misspecification by considering simulated data from a T mixture and a two-piece iskew-T mixture and illustrate computations under product of Binomial mixtures.

In Chapter 6 we propose a new computational strategy that provides a direct connection between cluster occupancies and Bayes factors with the advantage that Bayes factors allow for more general model comparisons (for instance equal vs unequal covariances in Normal mixtures). Likewise this algorithm offers a connection between posterior probabilities and empty cluster probabilities, hence we called it the ECP (empty cluster probability) estimator. In this chapter we also explore a

fast computational non-local model selection criteria.

In Chapter 7 we present the applications of this thesis. We consider the Old-Faithful, flow cytometry, Fisher's Iris and USA political blog data sets to illustrate the performance of our MOM-IW and MOM-Beta priors with respect to Normal-IW and Beta, BIC, AIC, sBIC. We also provide a comparison with overfitted and repulsive overfitted mixtures.

Conclusions and some possible future directions of this work are presented in Chapter 8.

In Appendix A we present the proofs. We illustrate in Appendix B the usage of diagnostics for MCMC runs of the considered examples. Appendix C presents the main probability density functions used throughout this thesis.

Our methodology is implemented in R packages mombf and NLPmix available at CRAN and `https://warwick.ac.uk/fac/sci/statistics/staff/research_students/patino_fuquene`, respectively.

Chapter 2

Theoretical framework

A NLP under $\mathcal{M}_k$ assigns vanishing density to any $\boldsymbol{\vartheta}_k$ such that (1.1.1) is equivalent to a mixture with $k' < k$ components. A necessary condition is to avoid vanishing ($\eta_j = 0$) and overlapping components ($\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$) but for this to also be a sufficient condition one needs to require *generic identifiability*. Definition 2 is adapted from Leroux (1992).

Definition 2 *Let* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \sum_{j=1}^{k} \eta_j p(\mathbf{y} \mid \boldsymbol{\theta}_j)$ *and* $p(\mathbf{y} \mid \widetilde{\boldsymbol{\vartheta}}_{\tilde{k}}, \mathcal{M}_{\tilde{k}}) = \sum_{j=1}^{\tilde{k}} \tilde{\eta}_j p(\mathbf{y} \mid \widetilde{\boldsymbol{\theta}}_j)$ *be two mixtures as in (1.1.1). Assume that* $\eta_j > 0, \tilde{\eta}_j > 0$ *for all* j *and that* $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_{j'}$, $\widetilde{\boldsymbol{\theta}}_j \neq \widetilde{\boldsymbol{\theta}}_{j'}$ *for all* $j \neq j'$. *The class* $p(\mathbf{y} \mid \boldsymbol{\theta})$ *defines a generically identifiable mixture if* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p(\mathbf{y} \mid \widetilde{\boldsymbol{\vartheta}}_{\tilde{k}}, \mathcal{M}_{\tilde{k}})$ *for almost every* $\mathbf{y}$ *implies that* $k = \tilde{k}$ *and* $\boldsymbol{\vartheta}_k = \tilde{\boldsymbol{\vartheta}}_{\Psi(\tilde{k})}$ *for some permutation* $\Psi(\tilde{k})$ *of the component labels in* $\mathcal{M}_{\tilde{k}}$.

That is, assuming that all components have non-zero weights and distinct parameters the mixture is uniquely identified by its parameters up to label permutations. Teicher (1963) showed that mixtures of univariate Normal, Exponential and Gamma distributions are generically identifiable. Yakowitz and Spragins (1968) extended the result to several multivariate distributions, including the Normal case. See also Allman et al. (2009) for a study of strong identifiability for multivariate Bernoulli mixtures, finite and infinite product Binomial mixtures, hidden Markov Models and random graph mixture models. In particular product Binomial mixtures are generically identifiable when the number of Binomial trials is above a small threshold (Allman et al. (2009), Theorem 4), e.g. when the number of trials $L_{if} = L$ for all (i, f) then it suffices that $3L^{p/3} > 2(k+1)$.

Throughout we assume $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ to be generically identifiable. Then $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ defines a NLP under Definition 1. In this chapter we present the theoretical framework of this research. In Section 2.1 we define a new general NLP

class for mixture distributions. In Section 2.2 we present the theoretical conditions that will be used in the proof of our main result (Theorem 1). In Section 2.3 we offer in Theorem 1 a theoretical characterization of the sparsity induced by NLPs to choose the number of components in a mixture.

2.1 A general NLP class for mixture distributions

Let $d_\vartheta(\boldsymbol{\vartheta}_k)$ be a continuous penalty function converging to 0 under (i) or (ii), then a general NLP class is defined by

$$p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k), \tag{2.1.1}$$

where $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ is an arbitrary LP with the restriction that $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ is proper. We consider $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = p^L(\boldsymbol{\theta} \mid \mathcal{M}_k) p^L(\boldsymbol{\eta} \mid \mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k) = d_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta})$, where

$$d_\theta(\boldsymbol{\theta}) = \frac{1}{C_k} \left(\prod_{1 \leq i < j \leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \right), \tag{2.1.2}$$

is a repulsive force between components akin to Petralia et al. (2012), $C_k = \int p^L(\boldsymbol{\theta} \mid \mathcal{M}_k) \prod_{1 \leq i < j \leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) d\boldsymbol{\theta}$ a prior normalization constant and $d_\eta(\boldsymbol{\eta}) \propto \prod_{i=1}^k \eta_j^r$ with $r > 0$. Evaluating C_k may require numerical approximations (e.g. Monte Carlo) but in the next section we give closed expressions for specific $d_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)$. Regarding the weights, we set the symmetric Dirichlet $p(\boldsymbol{\eta} \mid \mathcal{M}_k) = \text{Dir}(\boldsymbol{\eta}; q) \propto d_\eta(\boldsymbol{\eta}) \text{Dir}(\boldsymbol{\eta}; q - r)$, where importantly one must set $q > 1$ to satisfy (i) above and $r \in [q - 1, q)$. Summarizing, we set

$$p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\theta(\boldsymbol{\theta}) p^L(\boldsymbol{\theta} \mid \mathcal{M}_k) \text{Dir}(\boldsymbol{\eta}; q), \tag{2.1.3}$$

where $q > 1$ and $d_\theta(\boldsymbol{\theta})$ is as in (2.1.2).

The specific form of $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ depends on the model under consideration. For instance consider $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$ for a location parameter $\boldsymbol{\mu}_i$ and scale matrix Σ_i. Then one may adapt earlier proposals for variable selection and define MOM penalties (Johnson and Rossell, 2010) $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/g$ where A is a symmetric positive-definite matrix, or alternatively eMOM penalties (Rossell et al., 2013) $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp\{-g/(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}$ where g is a prior dispersion parameter, also adopted by Petralia et al. (2012) for repulsive mixtures. Note that C_k is guaranteed to be finite for eMOM penalties as $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \leq 1$. The main difference between MOM and eMOM is that the latter induce a stronger model

separation that give faster sparsity rates.

However, empirical results in Johnson and Rossell (2010, 2012) and Rossell and Telesca (2017) suggest that by setting g adequately both MOM and eMOM are often equally satisfactory.

Although our theory holds for fairly general $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, we now propose simple choices leading to convenient interpretation and closed-form C_k.

2.1.1 Application to Normal and T mixtures

We consider first the case where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$, $\boldsymbol{\mu}_i$ is a location parameter and Σ_i a positive-definite matrix, as in Normal or T mixtures. Then in (2.1.3) we may set the MOM-Inverse Wishart (MOM-IW) prior

$$p(\boldsymbol{\theta} \mid \mathcal{M}_k) = d_\theta(\boldsymbol{\theta}) p^L(\boldsymbol{\theta} \mid \mathcal{M}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \times \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j \mid \mathbf{0}, g A_\Sigma\right) \mathrm{IW}(\Sigma_j \mid \nu, S), \quad (2.1.4)$$

where A_Σ^{-1} is a symmetric positive-definite matrix and (g, ν, S) are fixed prior hyperparameters. A trivial choice is $A_\Sigma^{-1} = I$ but it has the inconvenience of not being invariant to changes in scale of $\mathbf{y}$. Instead we use $A_\Sigma^{-1} = \frac{1}{k} \sum_{j=1}^{k} \Sigma_j^{-1}$, which is symmetric and positive-definite and is related to the L_2 distance between Normal distributions. In the particular case where $\Sigma_1 = \ldots = \Sigma_k = \Sigma$, a parsimonious model sometimes considered to borrow information across components, clearly $A_\Sigma = \Sigma$. In our model-fitting algorithms and examples we consider both the equal and unequal covariance cases. We remark that in the latter case (2.1.4) defines a NLP that penalizes $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$ even when $\Sigma_i \neq \Sigma_j$. We do not view this as problematic, given that in most applications the interest is to identify components with well-separated locations. We note however that if one is interested in detecting components that differ only in $\Sigma_i \neq \Sigma_j$ then $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ should be adjusted. In general one may set $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ to any measure of distance or divergence between probability distributions. As illustration, one could use the squared Hellinger distance between Normal distributions

$$d_\theta(\boldsymbol{\theta}) = \frac{1}{C_k} \prod_{1\leq i<j\leq k} 1 - \frac{\det(\Sigma_i)^{1/4}\det(\Sigma_j)^{1/4}}{\det\left((\Sigma_i+\Sigma_j)/2\right)^{1/2}} \times \exp\left\{-\frac{1}{8}\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)'2(\Sigma_i+\Sigma_j)^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{g}\right\}. \tag{2.1.5}$$

For this choice $d_\theta(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$ and $\Sigma_i = \Sigma_j$. Alternatively we may consider eMOM penalties (Rossell et al., 2013) given by $d_\theta(\boldsymbol{\theta}) = \exp\{-g/(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}$, also adopted by Petralia et al. (2012) for repulsive mixtures.

In the univariate case (p=1) (2.1.4) can be written

$$p(\boldsymbol{\theta} \mid \mathcal{M}_k) = \frac{1}{C_k} \prod_{1\leq i<j\leq k} \frac{(\mu_i-\mu_j)' A_\Sigma^{-1}(\mu_i-\mu_j)}{g} \prod_{j=1}^{k} N\left(\mu_j \mid 0, gA_{\sigma^2}\right) \mathrm{IG}(\sigma_j^2 \mid \nu, S), \tag{2.1.6}$$

where ν and S are the hyperparameters of the Inverse-Gamma distribution and $A_{\sigma^2}^{-1} = \frac{1}{k}\sum_{j=1}^{k}(\sigma_j^2)^{-1}$. To illustrate consider a sample $y_1, ..., y_n$ of size n for testing one component versus a two-component univariate Normal mixture as follows

$$\mathcal{M}_1 : y_i \sim N(y_i; \mu, \sigma^2) \quad \text{vs} \quad \mathcal{M}_2 : y_i \sim \eta N(y_i; \mu_1, \sigma^2) + (1-\eta)N(y_i; \mu_2, \sigma^2),$$

where σ^2 and η are known and the prior probabilities of each model are $P(\mathcal{M}_1) = P(\mathcal{M}_2) = 1/2$. Under $\mathcal{M}_1$ the prior for μ is a Normal distribution given by

$$p(\mu|\sigma^2, m, g = 1, \mathcal{M}_1) = N(\mu; m, \sigma^2).$$

The Normal and Moment priors for the component means under $\mathcal{M}_2$ are respectively

$$p^L(\mu_1, \mu_2 \mid \sigma^2, m, g^L, \mathcal{M}_2) = N(\mu_1; m, \sigma^2 g^L)N(\mu_2; m, \sigma^2 g^L), \tag{2.1.7}$$

$$p(\mu_1, \mu_2 \mid \sigma^2, m, g, \mathcal{M}_2) = \frac{(\mu_2-\mu_1)^2}{2\sigma^2 g} N(\mu_1; m, \sigma^2 g)N(\mu_2; m, \sigma^2 g). \tag{2.1.8}$$

The top panels in Figure 2.1 illustrate how the Normal prior (right) assigns high prior density to $\mu_1 = \mu_2$, whereas the MOM prior incorporates a separation between $\mathcal{M}_1$ and $\mathcal{M}_2$. Notice that g is important for prior elicitation as it drives this separation (see Chapter 3). Consider the Normal and MOM priors using the separation parameters $\sqrt{\kappa} = (\mu_2 - \mu_1)/\sigma$ and $\mu_1^* = \mu_1/\sigma$ as follows:

$$p^L(\mu_1^*, \sqrt{\kappa} \mid \sigma^2, m, g^L, \mathcal{M}_2) = N(\mu_1^*; m, g^L)N(\sqrt{\kappa}; m - \mu_1^*, g^L);$$
$$p(\mu_1^*, \sqrt{\kappa} \mid \sigma^2, m, g, \mathcal{M}_2) = \frac{\kappa}{2g}N(\mu_1^*; m, g)N(\sqrt{\kappa}; m - \mu_1^*, g).$$

Figure 2.1 (Bottom) displays the MOM prior which induces a penalization term of $\kappa = (\mu_2 - \mu_1)^2/\sigma^2$ which is the natural unit of measure of separability between two clusters proposed by Fisher (1936).

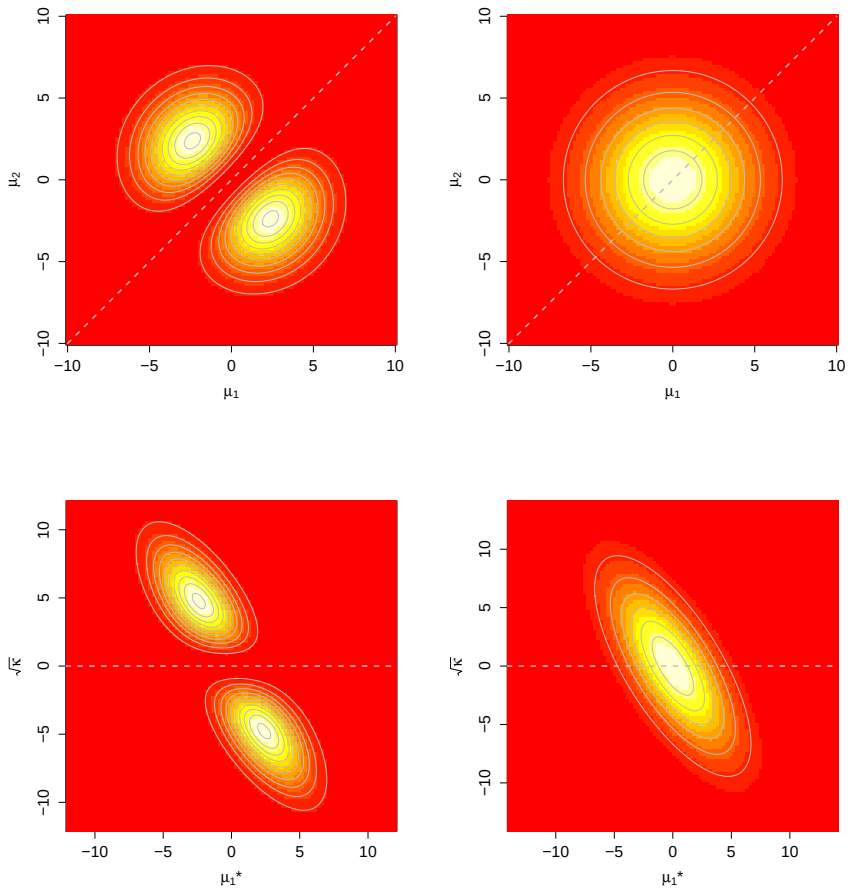


Figure 2.1: Top: Default MOM $p(\mu_1, \mu_2 \mid \sigma^2 = 1, m = 0, g = 5.68, \mathcal{M}_2)$ (left) and Normal $p^L(\mu_1, \mu_2 \mid \sigma^2 = 1, m = 0, g^L = 11.56, \mathcal{M}_2)$ (right). Bottom: Default MOM $p(\mu_1^*, \sqrt{\kappa} \mid \sigma^2 = 1, m = 0, g = 5.68, \mathcal{M}_2)$ (left) and Normal $p^L(\mu_1^*, \sqrt{\kappa} \mid \sigma^2 = 1, m = 0, g^L = 11.56, \mathcal{M}_2)$ (right).

2.1.2 Application to Binomial mixtures

We now consider binary data, specifically for a Binomial mixture the MOM-Beta prior

$$p(\boldsymbol{\theta} \mid \mathcal{M}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} (\theta_i - \theta_j)^2 \prod_{j=1}^{k} \text{Beta}(\theta_j; ag, (1-a)g), \tag{2.1.9}$$

where $\theta_j > 0$ is the success probability in component j and $a > 0$, $g > 0$ are known prior parameters. In our parameterization $a > 0$ is the prior mean and $g > 0$ the prior sample size for the underlying Beta prior. Figure 2.2 displays the implied prior density and for comparison, that for a Beta prior setting $g^L = 1.98$ to match the prior variance of the Beta$(1, 1)$ (see Chapter 3 for the prior elicitation of g).

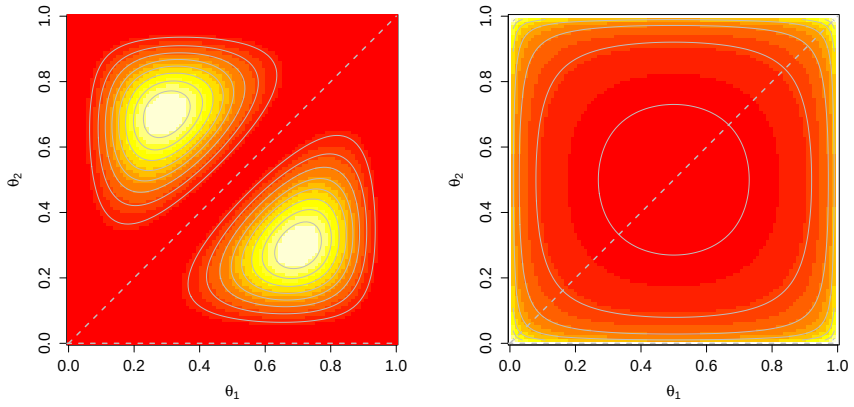


Figure 2.2: Default MOM-Beta $p(\boldsymbol{\theta} \mid g = 7.05, \mathcal{M}_2)$ (left) and Beta $p^L(\boldsymbol{\theta} \mid g^L = 1.98, \mathcal{M}_2)$ (right)

2.1.3 Application to product Binomial mixtures

For a product Binomial mixture (Binomial mixtures are the particular case $p = 1$) we define the MOM-Beta prior

$$p(\boldsymbol{\theta} \mid \mathcal{M}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \prod_{j=1}^{k} \prod_{f=1}^{p} \text{Beta}(\theta_{jf}; ag, (1-a)g), \tag{2.1.10}$$

where $\theta_{jf} > 0$ is the success probability for outcome f in component j and $a > 0$, $g > 0$ are known prior parameters. In our parameterization $a > 0$ is the prior mean and $g > 0$ the prior sample size for the underlying Beta prior.

We now turn our attention in Sections 2.2-2.3 to the theoretical results for the sparsity induced by the NLP described in equation (2.1.1) in the context of choosing the number components in a mixture.

2.2 Parsimony enforcement

We now offer some theoretical results for both MOM and eMOM penalties, but in our implementations we focus on the MOM for the practical reasons that C_k has closed form and leads to simple prior elicitation. Both MOM and eMOM remain applicable when $\boldsymbol{\theta}_i$ is a vector of probabilities, as we illustrate for Binomial and product Binomial mixtures. More generally $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ can be based on any distance or divergence between probability measures. We defer discussion of prior elicitation to Chapter 3.

We show that NLPs induce extra parsimony via the penalty term $d_\vartheta(\boldsymbol{\vartheta}_k)$, which affects specifically overfitted mixtures. We first lay out technical conditions for the result to hold. Recall that k^* is the true number of components and $\boldsymbol{\vartheta}^*_{k^*}$ the true parameter value. Let $p^*_k(\mathbf{y})$ be the density minimising Kullback-Leibler (KL) divergence between the data-generating $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$ and the class $\{p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k), \boldsymbol{\vartheta}_k \in \Theta_k\}$. When $k \leq k^*$ for generically identifiable mixtures $p^*_k(\mathbf{y})$ is defined by a unique parameter $\boldsymbol{\vartheta}^*_k \in \Theta_k$ (up to label permutations). When $k > k^*$ there are multiple minimizers giving $p^*_k(\mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$. $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ denotes a LP and $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ a NLP as in (2.1.1). $P^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ and $E^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ are the posterior probability and expectation under $p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$.

2.2.1 Technical conditions

B1 L_1 *consistency.* For all fixed $\epsilon > 0$ as $n \to \infty$

$$P^L\left(\int |p(\mathbf{z} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) - p^*_k(\mathbf{z})|\, d\mathbf{z} > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

in probability with respect to $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$.

B2 *Continuity.* $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ is a continuous function in $\boldsymbol{\vartheta}_k$.

B3 *Penalty boundedness.* There is a constant c_k such that $d_\vartheta(\boldsymbol{\vartheta}_k) \leq c_k$ for all $\boldsymbol{\vartheta}_k$. Alternatively, if $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ involves the MOM-IW prior (2.1.4) and $k > k^*$

then there exist finite $\epsilon, U > 0$ such that

$$\lim_{n\to\infty} P\left(E^L\left[\exp\left\{\frac{1}{2g}\sum_{j=1}^{k}\boldsymbol{\mu}_j' A_\Sigma^{-1}\boldsymbol{\mu}_j\frac{\epsilon}{1+\epsilon}\right\} \mid \mathbf{y}, \mathcal{M}_k\right] < U\right) = 1.$$

2.2.2 Additional conditions from Rousseau and Mengersen (2011)

We reproduce Conditions A1-A4 in Rousseau and Mengersen (2011), adjusted to the notation we used in this work. Their Condition A5 is trivially satisfied by our $\boldsymbol{\eta} \sim \text{Dir}(q)$ prior, hence is not reproduced here. Recall that we defined $p^*_{k^*}(\mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$ to be the data-generating truth.

We denote $\Theta^*_k = \{\boldsymbol{\vartheta}_k \in \Theta_k; p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p^*_{k^*}(\mathbf{y})\}$ and let $\log(p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k))$ be the log-likelihood calculated at $\boldsymbol{\vartheta}_k$. Denote $F_0(g) = \int p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})g(\mathbf{y})d\mathbf{y}$ where $g(\cdot)$ is a probability density function, denote by $\text{Leb}(A)$ the Lebesgue measure of a set A and let $\nabla p(\mathbf{y} \mid \boldsymbol{\theta})$ be the vector of derivatives of $p(\mathbf{y} \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and $\nabla^2 p(\mathbf{y} \mid \boldsymbol{\theta})$ be the second derivatives with respect to $\boldsymbol{\theta}$. Define for $\epsilon \geq 0$

$$\bar{p}(\mathbf{y} \mid \boldsymbol{\theta}) = \sup_{|\boldsymbol{\theta}^l - \boldsymbol{\theta}| \leq \epsilon} p(\mathbf{y} \mid \boldsymbol{\theta}^l), \qquad \underline{p}(\mathbf{y} \mid \boldsymbol{\theta}) = \inf_{|\boldsymbol{\theta}^l - \boldsymbol{\theta}| \leq \epsilon} p(\mathbf{y} \mid \boldsymbol{\theta}^l).$$

We now introduce some notation that is useful to characterize Θ^*_k, following Rousseau and Mengersen (2011). Let $\mathbf{w} = (w_i)_{i=0}^{k^*}$ with $0 = w_0 < w_1 < ... < w_{k^*} \leq k$ be a partition of $\{1, ..., k\}$. For all $\boldsymbol{\vartheta}_k \in \Theta_k$ such that $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p^*_k(\mathbf{y})$ there exists $\mathbf{w}$ as defined above such that, up to a permutation of the labels,

$$\forall i = 1, ..., k^*, \quad \boldsymbol{\theta}_{w_{i-1}+1} = ... = \boldsymbol{\theta}_{w_i} = \boldsymbol{\theta}^*_i, \quad \eta(i) = \sum_{j=w_{i-1}+1}^{w_i} \eta_j = \eta^*_i, \quad \eta_{w_{k^*+1}} = ... = \eta_k = 0.$$

In other words, $I_i = \{w_{i-1} + 1, ..., w_i\}$ represents the cluster of components in $\{1, ..., k\}$ having the same parameter as $\boldsymbol{\theta}^*_i$. Then define the following parameterisation of $\boldsymbol{\vartheta}_k \in \Theta_k$ (up to permutation)

$$\boldsymbol{\iota}_{\boldsymbol{w}} = \left((\boldsymbol{\theta}_j)_{j=1}^{w_{k^*}}, (r_i)_{i=1}^{k^*-1}, (\eta_j)_{j=w_{k^*+1}}^{k}\right) \in \mathbb{R}^{pw_{k^*}+k^*+k-w_{k^*}-1}, \quad r_i = \eta(i) - \eta^*_i, \quad i = 1, ..., k^*,$$

and

$$\boldsymbol{\varpi}_{\boldsymbol{w}} = \left((f_j)_{j=1}^{w_{k^*}}, \boldsymbol{\theta}_{w_{k^*}+1}, ..., \boldsymbol{\theta}_k\right), \quad f_j = \frac{\eta_j}{\eta(i)}, \quad \text{when} \quad j \in I_i = \{w_{i-1} + 1, ..., w_i\},$$

note that for $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$

$$\boldsymbol{\iota}^*_{\boldsymbol{w}} = (\boldsymbol{\theta}^*_1, ..., \boldsymbol{\theta}^*_1, \boldsymbol{\theta}^*_2, ..., \boldsymbol{\theta}^*_2, ..., \boldsymbol{\theta}^*_{k^*}, ..., \boldsymbol{\theta}^*_{k^*}, 0...0...0)$$

where $\boldsymbol{\theta}^*_i$ is repeated $w_i - w_{i-1}$ times in the above vector for any $\boldsymbol{\varpi}_{\boldsymbol{w}}$. Then we parameterize $(\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$, so that $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) = p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ and we denote $\nabla p(\mathbf{y} \mid (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ and $\nabla^2 p(\mathbf{y} \mid (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ the first and second derivatives of $p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)$ with respect to $\boldsymbol{\iota}_{\boldsymbol{w}}$ and computed at $\boldsymbol{\vartheta}^*_{k^*} = (\boldsymbol{\iota}^*_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$. We also denote by $P^L(\cdot \mid \mathbf{y}, \mathcal{M}_k)$ the posterior distribution using a LP.

Conditions

A1 L_1 *consistency.* For all $\epsilon = (\log n)^e/\sqrt{n}$ with $e \geq 0$ as $n \to \infty$

$$P^L\left(\int |p(\mathbf{z} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) - p^*_k(\mathbf{z})|\, d\mathbf{z} > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

in probability with respect to $p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*})$.

A2 *Regularity.* The component density $p(\mathbf{y} \mid \boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta} \in \Theta$ is three times differentiable and regular in the sense that for all $\boldsymbol{\theta} \in \Theta$ the Fisher information matrix associated with $p(\mathbf{y} \mid \boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta}$. Denote $\nabla^3 p(\mathbf{y} \mid \boldsymbol{\theta})$ the array whose components are

$$\frac{\partial^3 p(\mathbf{y} \mid \boldsymbol{\theta})}{\partial_{\boldsymbol{\theta}_{i1}} \partial_{\boldsymbol{\theta}_{i2}} \partial_{\boldsymbol{\theta}_{i3}}}$$

For all $i \leq k^*$, there exists $\epsilon > 0$ such that

$$F_0\left(\frac{\bar{p}(\mathbf{y} \mid \boldsymbol{\theta}^*_i)^3}{\underline{p}(\mathbf{y} \mid \boldsymbol{\theta}^*_i)^3}\right) < \infty, \quad F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*| \leq \epsilon} |\nabla p(\mathbf{y}|\boldsymbol{\theta})|^3}{\underline{p}(\mathbf{y} \mid \boldsymbol{\theta}^*_i)^3}\right) < \infty, \quad F_0\left(\frac{|p(\mathbf{y}|\boldsymbol{\theta}^*_i)|^4}{(p(\mathbf{y} \mid \boldsymbol{\vartheta}^*_{k^*}, \mathcal{M}_{k^*}))^4}\right) < \infty,$$

$$F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*| \leq \epsilon} |\nabla^2 p(\mathbf{y} \mid \boldsymbol{\theta})|^2}{\underline{p}(\mathbf{y} \mid \boldsymbol{\theta}^*_i)^2}\right) < \infty, \quad F_0\left(\frac{\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}^*| \leq \epsilon} |\nabla^3 p(\mathbf{y} \mid \boldsymbol{\theta})|^2}{\underline{p}(\mathbf{y} \mid \boldsymbol{\theta}^*_i)}\right) < \infty.$$

Assume also that for all $i = 1, ..., k^*$, $\boldsymbol{\theta}^*_i \in \text{int}(\Theta^k)$ the interior of Θ^k.

A3 *Integrability.* There exists $\Theta^{k^*} \subset \Theta^k$ satisfying $\text{Leb}(\Theta^{k^*}) > 0$ and for all $i \leq k^*$

$$d(\boldsymbol{\theta}_i^*, \Theta^{k^*}) = \inf_{\boldsymbol{\theta} \in \Theta^{k^*}} |\boldsymbol{\theta} - \boldsymbol{\theta}_i^*| > 0$$

and such that for all $\boldsymbol{\theta} \in \Theta^{k^*}$,

$$F_0\left(\frac{p(\mathbf{y} \mid \boldsymbol{\theta})^4}{(p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*}))^4}\right) < \infty, \qquad F_0\left(\frac{p(\mathbf{y} \mid \boldsymbol{\theta})^3}{\underline{p}(\mathbf{y} \mid \boldsymbol{\theta}_i^*)^3}\right) < \infty, \qquad \forall i \leq k^*.$$

A4 *Stronger identifiability.*

For all $\mathbf{w}$ partitions of $\{1, ..., k\}$ as defined above, let $\boldsymbol{\vartheta}_k \in \Theta_k$ and write $\boldsymbol{\vartheta}_k$ as $(\boldsymbol{\iota}_{\boldsymbol{w}}, \boldsymbol{\varpi}_{\boldsymbol{w}})$; then

$$(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*)' \nabla p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}^*, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k) + \frac{1}{2}(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*)' \nabla^2 p(\mathbf{y} \mid (\boldsymbol{\iota}_{\boldsymbol{w}}^*, \boldsymbol{\varpi}_{\boldsymbol{w}}), \mathcal{M}_k)(\boldsymbol{\iota}_{\boldsymbol{w}} - \boldsymbol{\iota}_{\boldsymbol{w}}^*) = 0 \Leftrightarrow$$
$$\forall i \leq k^*, r_i = 0 \text{ and } \forall j \in I_i \;\; f_j(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*) = 0, \;\; \forall i \geq w_{k^*} + 1, \;\; p_i = 0.$$

Assuming also that if $\boldsymbol{\theta} \notin \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k\}$ then for all functions $h_{\boldsymbol{\theta}}$ which are linear combinations of derivatives of $p(\mathbf{y} \mid \boldsymbol{\theta})$ of order less than or equal to 2 with respect to $\boldsymbol{\theta}$, and all functions h_1 which are also linear combinations of derivatives of the $p(\mathbf{y} \mid \boldsymbol{\theta}_j)$'s $j = 1, 2, .., k$ and its derivatives of order less or equal to 2, then $\alpha h_{\boldsymbol{\theta}} + \beta h_1 = 0$ if and only if $\alpha h_{\boldsymbol{\theta}} = \beta h_1 = 0$.
Extension to non compact cases: If Θ^k is not compact then we also assume that for all sequences $\boldsymbol{\theta}_n$ converging to a point in $\partial\Theta^k$ the frontier of Θ^k, considered as a subset of $\Re \cup \{-\infty, \infty\}^p$, $p(\mathbf{y} \mid \boldsymbol{\theta}_n)$ converges pointwise either to a degenerate function or to a proper density $p(\cdot)$ such that $p(\cdot)$ is linearly independent of any null combinations of $p^*(\mathbf{y} \mid \boldsymbol{\theta}_i)$, $\nabla p^*(\mathbf{y} \mid \boldsymbol{\theta}_i)$ and $\nabla^2 p^*(\mathbf{y} \mid \boldsymbol{\theta}_i)$, $i = 1, ..., k^*$.

2.2.3 Discussion of the technical conditions

Condition B1 amounts to posterior L_1 consistency of $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ to the data-generating truth when $k \geq k^*$ and to the KL-optimal density when $k < k^*$. Note that B1 is assumed under the underlying local p^L and hence follows from standard theory. Specifically, B1 is a milder version of Condition A1 in Rousseau and Mengersen (2011) where rather than fixed ϵ one has $\epsilon = \sqrt{\log n}/\sqrt{n}$. See the discus-

sion therein and Ghosal and der Vaart (2001) for results on finite Normal mixtures, Rousseau (2007) for Beta mixtures and Ghosal and Van Der Vaart (2007) for infinite Normal mixtures. For strictly positive $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) > 0$ Condition B1 is intimately connected to MLE consistency (Ghosal, 2002), proven for fairly general mixtures by Redner (1981) for $k \leq k^*$ and by Leroux (1992) for $k > k^*$.

The L_1 consistency results above focus on the case where the data-generating truth lies in the assumed family, but see Ramamoorthi et al. (2015) (Theorem 2) for posterior concentration results under model misspecification for independent and identically distributed data. Condition B2 holds when the kernel $p(\mathbf{y} \mid \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, as in the vast majority of common models. B3 is trivially satisfied when NLPs are defined using bounded penalties. For the MOM-IW we require the technical condition that the posterior exponential moment in B3 is bounded in probability when $k > k^*$ (Lemma A.1.1). To gain intuition, B3 requires that under the posterior distribution $p^L(\boldsymbol{\mu} \mid \mathcal{M}_k, \mathbf{y})$ none of the elements in $\boldsymbol{\mu}$ diverges to infinity, and in particular is satisfied if $\boldsymbol{\mu}$ is restricted to a compact support.

2.3 Theoretical characterization of the sparsity

Theorem 1 below states that $d_\vartheta(\boldsymbol{\vartheta}_k)$ imposes a complexity penalty concentrating on 0 when $k > k^*$ and a constant when $k \leq k^*$. Part (i) applies to any model, Part (ii) only requires B1-B3 and Part (iii) holds under the mild conditions A1-A4 in Rousseau and Mengersen (2011) (Section 2.2.2), hence the result applies to an ample class of mixtures. The proof of Part(iii) only requires posterior contraction of the sum of redundant weights at a $n^{-1/2}$ rate, and can be trivially adjusted when this rate is slower. Rousseau and Mengersen (2011) showed that the $n^{-1/2}$ rate is achieved under Conditions A1-A3 and a strong identifiability condition A4. Interestingly, Ho and Nguyen (2016) showed that strong identifiability can be expressed in terms of partial differential equations involving the kernel $p(\mathbf{y} \mid \boldsymbol{\theta})$, its first and second derivatives. In particular location-scale Gaussian and Gamma mixtures are not strongly identifiable for certain problematic $\boldsymbol{\vartheta}_k$. When the data-generating $\boldsymbol{\vartheta}_k^*$ is one of those problematic values then the MLE of the component parameters $\hat{\boldsymbol{\theta}}$ is slower than $n^{-1/2}$, however remarkably the MLE of the mixing weights $\hat{\boldsymbol{\eta}}$ does contract at the $n^{-1/2}$ rate required by Part(iii).

Theorem 1 *Let $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ be a generically identifiable mixture, $p(\mathbf{y} \mid \mathcal{M}_k)$ and $p^L(\mathbf{y} \mid \mathcal{M}_k)$ the integrated likelihoods under $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ and $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$. Then*

(i) $p(\mathbf{y} \mid \mathcal{M}_k) = p^L(\mathbf{y} \mid \mathcal{M}_k) E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}\right)$, *where*

$$E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}\right) = \int d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k.$$

(ii) If B1-B2 are satisfied then as $n \to \infty$

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

where $d_k^* = 0$ *for* $k > k^*$ *and* $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*)$ *for* $k \leq k^*$.
If B3 also holds then $E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}\right) \xrightarrow{P} d_k^*$.

(iii) Let $k > k^*$ *and* $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) \propto d_\theta(\boldsymbol{\theta}) p^L(\boldsymbol{\theta} \mid \mathcal{M}_k) Dir(\boldsymbol{\eta}; q)$, *where* $q > 1$. *If B3 and A1-A4 in Rousseau and Mengersen (2011) hold for* $p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)$ *then for all* $\epsilon > 0$ *and all* $\delta \in (0, dim(\Theta)/2)$ *there exists a finite* $\tilde{c}_k > 0$ *such that*

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > \tilde{c}_k n^{-\frac{k-k^*}{2}(q-\delta)+\epsilon} \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

in probability as $n \to \infty$.

Part (i) extends Theorem 1 in Rossell and Telesca (2017) to mixtures and shows that $p(\mathbf{y} \mid \mathcal{M}_k)$ differs from $p^L(\mathbf{y} \mid \mathcal{M}_k)$ by a term $E^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}\right)$ that intuitively should converge to 0 for overfitted models. Part (i) also eases computation as $E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y})$ can be estimated from standard MCMC output from $p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$, as we exploit in Chapter 4. Part (ii) confirms that the posterior of $d_\vartheta(\boldsymbol{\vartheta}_k)$ under $p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ concentrates around 0 for overfitted models and a finite constant otherwise, and that its expectation also converges. Part (iii) states that for overfitted models this concentration rate is essentially $n^{-(k-k^*)q/2}$, leading to an accelerated sparsity-inducing Bayes factor $B_{k,k^*}(\mathbf{y}) = E^L(O_p(n^{-(k-k^*)q/2}))B^L_{k,k^*}(\mathbf{y})$ (See the proofs in Appendix A, Sections A.1 and A.2). Recall that as discussed earlier the LP-based $B^L_{k,k^*}(\mathbf{y}) = O_p(n^{-(\lambda - p_{k^*}/2)})$ for some $\lambda \in [p_{k^*}/2, p_k/2]$ under the conditions in Watanabe (2013). For instance, one might set q such that $(k-k^*)q/2 = \lambda - p_{k^*}/2$ so that $B_{k,k^*}(\mathbf{y})$ converges to 0 at twice the rate for $B^L_{k,k^*}(\mathbf{y})$. As λ is unknown in general one could conservatively take its upper bound $\lambda = p_k/2$, then $q = (p_k - p_{k^*})/(k - k^*)$ is the number of parameters per component. We further discuss prior elicitation in Chapter 3.

Chapter 3

Prior computation and elicitation

In this chapter we develop methodology required for the proposed framework to be practical. In Section 3.1 the challenging computation of the normalization constant for MOM priors is addressed, and in Section 3.2 the proposal for the prior elicitation is presented.

3.1 Prior normalization constant

We now discuss the computation of C_k in (2.1.4) and (2.1.10). This task requires a non-trivial expectation of a product of quadratic forms. Lemma 1 gives a recursive formula for C_k for any NLP with the generic form

$$p(\boldsymbol{\zeta} \mid \mathcal{M}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} (\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j)'(\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j) \prod_{j=1}^{k} \prod_{f=1}^{p} p^L(\boldsymbol{\zeta}_{jf}) \tag{3.1.1}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_k) \in \mathbb{R}^{pk}$. Note that (2.1.4) is the particular case where $\boldsymbol{\zeta}_i = A_{\Sigma}^{-1/2} \boldsymbol{\mu}_i$, and(2.1.10) is the case where $\boldsymbol{\zeta}_i = \boldsymbol{\theta}_i$.

Lemma 1 *Let $p(\boldsymbol{\zeta} \mid \mathcal{M}_k)$ be as in (3.1.1). Then*

$$C_k = \sum_{s \in S_k} \left(\prod_{l=1}^{pk} \kappa_{sl} \right) \sum_{v(1,2)=0}^{1} \cdots \sum_{v(k-1,k)=0}^{1} (-1)^{\sum_{i<j} v(i,j)} \left(\prod_{l=1}^{pk} \prod_{m=1}^{pk} \frac{b_{lm}^{s_{l,m}}(v)}{s_{l,m}!} \right)$$

where $\kappa_{sl} = E^L(\boldsymbol{\zeta}_{jf}^{\sum_{m=1}^{pk} s_{lm} + s_{ml}})$, $S_k = \left\{ (s_{1,1}, s_{1,2}, \ldots, s_{pk,pk}) : \sum_{l=1}^{pk} \sum_{m=1}^{pk} s_{l,m} = k(k-1)/2 \right\}$

with non-negative integers $0 \leq s_{l,m} \leq k(k-1)/2$*, and* $b_{lm}(v)$ *is the* (l,m) *element of the* $pk \times pk$ *matrix* B_v *given by*

$$\begin{cases} b_{ll} = \frac{1}{2}(k-1) - \sum_{i<j} v(i,j),\ l = 1 + p(i-1), \ldots, pi \\ b_{lm} = b_{ml} = -\frac{1}{2} + \sum_{i<j} v(i,j),\ (1+p(i-1), 1+p(j-1)), \ldots, (pi, pj) \end{cases}.$$

Proof. See Appendix A, Section A.3.

To illustrate Lemma 1 consider $p = 1$ and $k = 2$ and the normalization constant given by

$$C_k = \sum_{s \in S_k} \left(\prod_{l=1}^{2} \kappa_{sl} \right) \sum_{v(1,2)=0}^{1} (-1)^{v(1,2)} \left(\prod_{l=1}^{2} \prod_{m=1}^{2} \frac{b_{lm}^{s_{l,m}}(v)}{s_{l,m}!} \right)$$

where $S_k = \left\{ (s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}) : \sum_{l=1}^{2} \sum_{m=1}^{2} s_{l,m} = 1 \right\}$ with $0 \leq s_{l,m} \leq 1$ and $b_{11} = b_{22} = \frac{1}{2} - v(1,2)$, $b_{12} = b_{21} = -b_{11}$. We remark that Lemma 1 holds for any $p^L(\boldsymbol{\zeta})$ composed by independent and identically-distributed $p^L(\zeta_{jf})$ and that κ_s requires raw moments up to order $k(k-1)/2$, which can be pre-computed. For the MOM-Beta prior

$$\kappa_{sl} = \left(\frac{\Gamma(a+b)}{\Gamma(a)} \right)^{pk} \frac{\Gamma\left(a + \sum_{m=1}^{pk} s_{lm} + s_{ml} \right)}{\Gamma\left(a + b + \sum_{m=1}^{pk} s_{lm} + s_{ml} \right)}.$$

When p^L is a Normal prior the expression in Lemma 1 can be simplified, see Corollary 1. Further simplifications are possible when $p = 1$ or $k = 2$, these are given for Normal and product Binomial mixtures in Corollaries 2 and 3 respectively.

Corollary 1 *MOM-IW, general* (p,k)*. The normalization constant in (2.1.4) is*

$$C_k = \frac{1}{s!} \sum_{\upsilon_{(1,2)}=0}^{1} \ldots \sum_{\upsilon_{(k-1,k)}=0}^{1} (-1)^{\sum_{i,j}^{s} \upsilon_{(i,j)}} \mathcal{Q}_s(B_\upsilon), \tag{3.1.2}$$

where $v_{(i,j)} \in \{0,1\}$*,* $s = \binom{k}{2}$*,* $\mathcal{Q}_s(B_\upsilon) = s! 2^s d_s(B_\upsilon)$*,* $d_s(B_\upsilon) = \frac{1}{2s} \sum_{i=1}^{s} tr(B_\upsilon^i) d_{s-i}(B_\upsilon)$*,* $d_0(B_\upsilon) = 1$ *and* B_υ *is a* $pk \times pk$ *matrix with element* (l,m) *given by*

$$\begin{cases} b_{ll} = \dfrac{1}{2}(k-1) - \sum_{i<j} \upsilon_{(i,j)}, \quad l = 1 + p(i-1), \ldots, pi \\ b_{lm} = b_{ml} = -\dfrac{1}{2} + \sum_{i<j} \upsilon_{(i,j)}, \quad (l,m) = (1+p(i-1), 1+p(j-1)), \ldots, (pi, pj) \end{cases}$$

where $i \neq j$, $i = 1, \ldots, k$, $j = 1, \ldots, k$ *and* $b_{lm} = 0$ *otherwise.*

Proof. See Appendix A, Section A.4.

Corollary 2 *MOM-IW, univariate or two-component mixtures. Let* C_k *be as in (2.1.4)*

(i) If $p = 1$, *then* $C_k = \prod_{j=1}^{k} \Gamma(j+1)$.

(ii) If $k = 2$, *then* $C_k = 2p$.

Proof. See Appendix A, Section A.5.

Corollary 3 *MOM-Beta, univariate or two-component mixtures. Let* C_k *be as in (2.1.10)*

(i) If $p = 1$, *then* $C_k = \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^k \prod_{j=1}^{k} \frac{\Gamma(a+k-j)\Gamma(b+k-j)\Gamma(j+1)}{\Gamma(a+b+2k-j-1)}$.

(ii) If $k = 2$, *then* $C_k = 2p\frac{ab}{(a+b)^2(a+b+1)}$.

Proof. See Appendix A, Section A.6.

Despite having closed-form C_k its evaluation for general (p, k) can be cumbersome, e.g. S_k in Lemma 1 is the set of partitions of $k(k-1)/2$ and has size exponential in k (Andrews, 1998). The sum in (3.1.2) is simpler but contains $k(k-1)/2$ terms, still prohibitive for large k. A practical option for large k is to evaluate C_k via Monte Carlo as the prior mean of $d_k(\boldsymbol{\theta})$ under p^L and tabulate it upfront, prior to data analysis. This is particularly convenient in Corollary 1 where C_k does not depend on the prior parameter g. To facilitate the implementation of our methodology Tables 3.1-3.2 provide C_k for (2.1.4) and (2.1.10) (respectively) and various (p, k). In Tables 3.1-3.2 we have exact values for C_k when $p = 1$ or $k = 2$ computed using Corollaries 2 and 3.

3.2 Prior elicitation

A critical aspect in a NLP is its induced separation between components, driven by g and q in (2.1.3). We propose defaults that can be used in the absence of a priori knowledge, whenever the latter is available we naturally recommend to include it in the prior.

We start by discussing g, first for Normal and T mixtures and subsequently for Binomial and product Binomial mixtures. The main idea is that we wish to find

Table 3.1: Estimation of $\log(C_k)$ and associated standard error (se) via Monte Carlo for the MOM-IW prior where $k = 2, ..., 10$ and $p = 1, ..., 10$.

					p					
	1		2		3		4		5	
k	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se
2	0.69	0	1.39	0	1.79	0	2.08	0	2.30	0
3	2.49	0	4.57	<0.01	5.70	<0.01	6.51	<0.01	7.14	<0.01
4	5.66	0	9.83	<0.01	11.98	<0.01	13.51	<0.01	14.70	<0.01
5	10.45	0	17.36	<0.01	20.83	<0.01	23.25	<0.01	25.16	<0.01
6	17.03	0	27.27	0.04	32.26	0.02	35.99	0.03	38.58	<0.01
7	25.55	0	38.81	0.07	46.33	0.04	51.11	0.02	55.01	0.02
8	36.16	0	53.01	0.10	62.05	0.05	69.70	0.07	74.51	0.04
9	48.96	0	66.46	0.08	80.73	0.11	89.83	0.08	96.35	0.05
10	64.07	0	82.71	0.10	100.43	0.08	111.81	0.09	120.87	0.10
					p					
	6		7		8		9		10	
k	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se
2	2.48	0	2.64	0	2.77	0	2.89	0	3.00	0
3	7.66	<0.01	8.09	<0.01	8.48	<0.01	8.82	<0.01	9.12	<0.01
4	15.68	<0.01	16.51	<0.01	17.25	<0.01	17.90	<0.01	18.49	<0.01
5	26.72	<0.01	28.04	<0.01	29.23	<0.01	30.26	<0.01	31.22	<0.01
6	40.81	<0.01	42.78	0.01	44.47	<0.01	45.99	<0.01	47.35	<0.01
7	58.21	0.04	60.78	0.02	63.05	0.01	65.15	0.01	67.08	0.01
8	78.44	0.04	82.13	0.04	84.96	0.02	88.01	0.04	90.19	0.02
9	101.82	0.05	106.15	0.05	110.12	0.05	113.81	0.04	116.87	0.03
10	127.88	0.07	133.19	0.05	138.22	0.05	143.08	0.06	146.70	0.04

clearly-separated components, then one can interpret the data-generating process in terms of distinct sub-populations. We thus set g such that there is small prior probability that any two components are poorly-separated, that is giving rise to a unimodal density. In Normal mixtures the number of modes depends on non-trivial parameter combinations (Ray and Lindsay, 2005), but when $\eta_1 = \eta_2 = 0.5$ and $\Sigma_1 = \Sigma_2$ the mixture is bimodal when $\kappa = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 4$. Thus we set g such that $P(\kappa < 4 \mid \mathcal{M}_2) = 0.1$ or 0.05, say. This is trivial, the prior on κ implied by (2.1.4) is $p(\kappa \mid \mathcal{M}_2) = \text{Gamma}(\kappa; p/2 + 1, 1/(4g))$. For instance, in a univariate Normal mixture $g = 5.68$ gives $P(\kappa < 4 \mid \mathcal{M}_2) = 0.05$, Figure 2.1 (left) portrays the associated prior. For comparison the right panel shows a Normal prior with $g^L = 11.56$, which also assigns $P^L(\kappa < 4 \mid \mathcal{M}_2) = 0.05$. Based on simulation and sensitivity analyses (Section 5.4.2) we found $P(\kappa < 4 \mid \mathcal{M}_2) = 0.05$ to be slightly

Table 3.2: Estimation of $\log(C_k)$ with the standard error (se) via Monte Carlo for the MOM-Beta prior where $k = 2, ..., 10$ and $p = 1, ..., 10$.

	p									
	1		2		3		4		5	
k	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se
2	-2.78	0	-1.69	0	-1.11	0	-0.73	0	-0.43	0
3	-8.29	0	-4.89	<0.01	-3.17	<0.01	-2.08	<0.01	-1.20	<0.01
4	-16.50	0	-9.48	<0.01	-6.09	<0.01	-3.94	<0.01	-2.22	<0.01
5	-27.43	0	-15.40	0.01	-9.76	<0.01	-6.23	<0.01	-3.41	<0.01
6	-41.06	0	-22.66	0.02	-14.10	0.02	-8.89	0.01	-4.71	0.01
7	-57.39	0	-31.23	0.05	-19.22	0.04	-11.82	0.03	-6.09	0.01
8	-76.44	0	-41.24	0.09	-24.98	0.07	-15.13	0.06	-7.51	0.04
9	-98.19	0	-52.57	0.16	-31.40	0.16	-18.84	0.10	-9.03	0.07
10	-122.66	0	-65.99	0.21	-39.17	0.17	-23.31	0.12	-10.92	0.12
	p									
	6		7		8		9		10	
k	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se	$\log(C_k)$	se
2	-0.22	0	-0.04	0	0.12	0	0.25	0	0.38	0
3	-0.57	<0.01	-0.05	<0.01	0.43	<0.01	0.82	<0.01	1.19	<0.01
4	-0.97	<0.01	0.05	<0.01	1.00	<0.01	1.76	<0.01	2.48	<0.01
5	-1.36	<0.01	0.31	<0.01	1.87	<0.01	3.13	<0.01	4.30	<0.01
6	-1.68	<0.01	0.79	<0.01	3.09	<0.01	4.95	<0.01	6.69	<0.01
7	-1.87	0.02	1.52	0.01	4.71	0.01	7.29	0.01	9.69	<0.01
8	-1.98	0.02	2.54	0.03	6.74	0.01	10.16	0.01	13.32	0.01
9	-2.03	0.05	4.03	0.06	9.28	0.04	13.63	0.05	17.65	0.03
10	-1.65	0.11	5.39	0.07	12.19	0.08	17.73	0.09	22.65	0.07

preferable to 0.1 for balancing parsimony vs. sensitivity.

Regarding T mixtures, Došlá (2009) showed that a univariate mixture with two components and equal degrees of freedom υ is bimodal if $\kappa > 4\upsilon/(\upsilon + 2)$. More generally for multivariate T mixtures with equal Σ (Ray and Lindsay (2005), Theorem 1 and Remark 4) showed that all its modes lie in $\mathbf{y}(a) = a\boldsymbol{\mu}_1 + (1 - a)\boldsymbol{\mu}_2$ where $a \in [0, 1]$. It is easy to show that when $\eta_1 = \eta_2 = 0.5$ there is a unique minimum at $a = 1/2$ if and only if $\kappa > 4\upsilon/(\upsilon + p + 1)$, and then the mixture density is bimodal. This matches the result from Došlá (2009) for $p = 1$ and for Normal mixtures in Ray and Lindsay (2005) as $\upsilon \to \infty$. Summarising, by default we set g such that $P(\kappa < 4\upsilon/(\upsilon + p + 1) \mid \upsilon, \mathcal{M}_k) = 0.05$, where recall that $p(\kappa \mid \upsilon, \mathcal{M}_2) = \text{Gamma}(\kappa; p/2 + 1, 1/(4g))$. Note that to complete the prior one must also consider a prior the degrees of freedom υ, since this is a standard problem

we refer the reader to Rossell and Steel (2017), for a review of possible strategies. Note also that other strategies to set g arise from using other measures of separation, *e.g.* within/between sums of squares instead of unimodality (Malsiner-Walli et al., 2017), but we do not pursue this here.

Consider now the MOM-Beta prior (2.1.10). In contrast to continuous mixtures, here one cannot use multi-modality to set the prior parameters (a,g). Instead we specify (a,g) such that the amount of prior information is comparable to $\theta_{jl} \sim \text{Beta}(1,1)$, a prior commonly viewed as minimally informative. Specifically, under independent $\theta_{jl} \sim \text{Beta}(1,1)$ the variance of $\sum_{j=1}^{p} \theta_{2j} - \theta_{1j}$ is $p/6$, hence in the MOM-Beta prior we set (a,g) such that $\text{Var}(\sum_{j=1}^{p} \theta_{2j} - \theta_{1j}) = p/6$. For Binomial mixtures this results in $(a, g) = (1/2, 7.05)$, and to assess sensitivity we also considered $g = 16.09$ and $g = 29.99$, which imply $\text{Var}(\sum_{j=1}^{p} \theta_{2j} - \theta_{1j}) = p/12$ and $p/24$, respectively. In our experiments these g values yield a competitive performance but $g = 7.05$ was preferable (Chapter 5, Section 5.4.3).

To illustrate the choice of q, we consider under $\mathcal{M}_2$ three different possible values of q for the prior on η given by Dir(η;1/2,1/2), Dir(η;1,1) and Dir(η;3,3). Figure 3.1 illustrates how the Dir(η;1/2,1/2) prior does not place mass in the boundaries of $\mathcal{M}_2$, and places substantial mass in neighborhoods around 0 or 1 in contrast to the Dir$(\eta; 3, 3)$ prior. As a result, the Dir$(\eta; 3, 3)$ prior may produce more shrinkage toward 0 or 1.

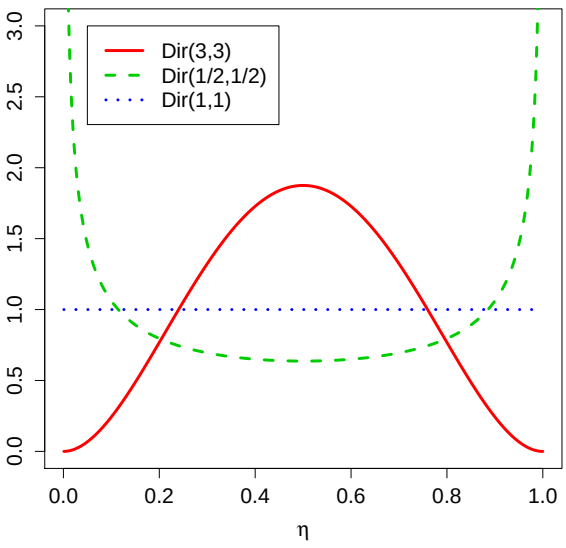


Figure 3.1: Illustration of prior densities for η.

Figure 3.1 shows how the uniform prior is equal to 1 for any value of η therefore this prior not only does not penalize the boundaries of $\mathcal{M}_2$, but also it does not produce any shrinkage around 0 or 1. Regarding q, as discussed earlier, $q > 1$ is required for (2.1.3) to define a NLP. One option is to set $q = 3$ so that $p(\boldsymbol{\eta} \mid \mathcal{M}_k) \propto \prod_{j=1}^{k} \eta_j^2$ induces a quadratic penalty comparable to the MOM prior on $\boldsymbol{\mu}$ given in (2.1.4). Alternatively from the discussion after Proposition 1 setting $q = (p_k - p_{k^*})/(k - k^*)$, the number of parameters per component, seeks to (at least) double the Bayes factor sparsity rate of the underlying LP. For instance, for Normal mixtures with common covariances this leads to $q = p + 1$, and under unequal covariances to $q = p + 0.5p(p+1) + 1$. These are the values we used in our examples with $p = 1$ or $p = 2$ (see Chapter 5), but we remark that for larger p such q may lead to an overly informative prior on $\boldsymbol{\eta}$ (see Figure 3.2). In our experience $q \in [2, 4]$ gives fairly robust results and satisfactory sparsity, thus larger values do not seem warranted.

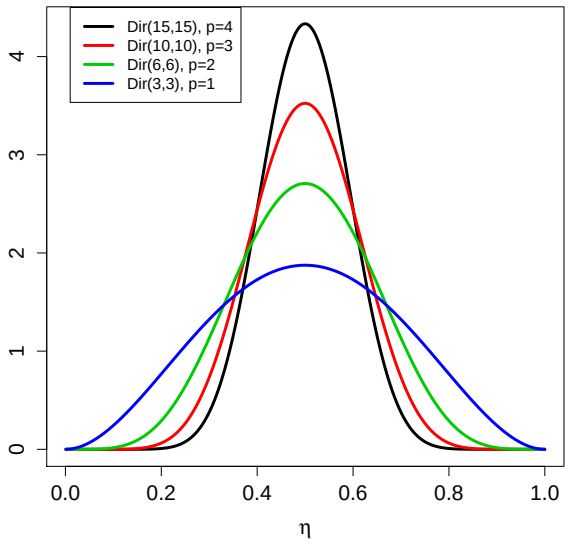


Figure 3.2: Illustration of prior densities for η with $p = \{1, 2, 3, 4\}$ for Normal mixtures with unequal covariances.

The prior distribution on the remaining parameters, which may be thought of as nuisance parameters, will typically reduce to a standard form for which defaults are already available. We assume that variables in the observed data are standardized to have mean 0 and variance 1 and set a default $S = (p + 4)^{-1} I$ and

$\nu = p + 4$, so that $E(\Sigma_j^{-1}) = I$. As an illustration for multivariate Normal mixtures we set $p(\Sigma_1, \ldots, \Sigma_k \mid \mathcal{M}_k) = \prod_{j=1}^{k} \mathrm{IW}(\Sigma_j; \nu, S)$. We follow the recommendation in Hathaway (1985) that eigenvalues of $\Sigma_i \Sigma_j^{-1}$ for any $i \neq j$ should be bounded away from 0 to prevent the posterior from becoming unbounded, which is achieved if $\nu \geq p + 4$ (Frühwirth-Schnatter (2006), Chapter 6). We remark that our framework can be sensitive to prior specification but, as we illustrated in this Chapter, default parameters based on multi-modality and minimal informativeness may represent a natural alternative for NLPs that result in a fairly competitive behaviour as we will explore in the next Chapters.

Chapter 4

Computational framework

Computation for mixtures is challenging, and potentially more so when embarking upon non-standard formulations such as ours. Fortunately, capitalising on Theorem 1(i) (Chapter 2) it is possible to estimate the integrated likelihood $p(\mathbf{y} \mid \mathcal{M}_k)$ for arbitrary mixtures with direct extensions of existing algorithms. Intuitively, one can use any algorithm to estimate the local prior integrated likelihood $p^L(\mathbf{y} \mid \mathcal{M}_k)$ and the posterior mean $E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}, \mathcal{M}_k)$.

In this chapter we present the computational framework of this thesis. In Section 4.1 we outline an MCMC algorithm to compute the integrated likelihood under MOM priors that only requires an MCMC run from the local posterior $p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ and is hence straightforward to implement using current software. We remark however that obtaining $p^L(\mathbf{y} \mid \mathcal{M}_k)$ can be costly and is the subject of current research (see Lee and Robert (2016) for a recent discussion). To estimate $p^L(\mathbf{y} \mid \mathcal{M}_k)$ we use the estimator proposed by Marin and Robert (2008). We found this estimator to be reasonably accurate, but it is limited to conjugate models and requires an MCMC post-processing step that may have non-negligible cost. Therefore in Chapter 6, Section 6.2 we propose a new estimator that only requires cluster probabilities available as an MCMC by-product, avoiding costly post-processing. Although our main interest is to infer k, in Section 4.2 we discuss posterior mode parameter estimates via an Expectation-Maximimation (EM) algorithm (Dempster et al. (1977)). Relative to $p^L(\mathbf{y} \mid \mathcal{M}_k)$, our $p(\mathbf{y} \mid \mathcal{M}_k)$ only requires a trivial MCMC post-processing step and the EM algorithm an extra gradient evaluation; both operations add a negligible cost relative to the corresponding LP calculations. In Section 4.3 we show comparisons of the proposed computational methods with existing approaches.

4.1 Integrated likelihood

Theorem 1(i) suggests the estimator

$$\hat{p}(\mathbf{y} \mid \mathcal{M}_k) = \tilde{p}(\mathbf{y} \mid \mathcal{M}_k)\frac{1}{T}\sum_{t=1}^{T}\omega(\boldsymbol{\vartheta}_k^{(t)}), \tag{4.1.1}$$

where $\omega(\boldsymbol{\vartheta}_k) = p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)/\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ and $\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ is an arbitrary LP conveniently chosen so that MCMC algorithms to sample $\boldsymbol{\vartheta}_k^{(t)} \sim \tilde{p}(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) \propto p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ are readily available. We remark that $\omega(\boldsymbol{\vartheta}_k)$ is not a reweighting to convert samples from $\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ into samples from $p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$, but a direct approximation to the posterior mean of $d(\boldsymbol{\vartheta}_k)$ under $\tilde{p}(\boldsymbol{\vartheta} \mid \mathbf{y}, \mathcal{M}_k)$. However, if interested in posterior samples from $p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ one could clearlyuse such a reweighting. For the MOM-IW in (2.1.4) we used

$$\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta}; q)\prod_{j=1}^{k} N(\boldsymbol{\mu}_j; \mathbf{0}, g\Sigma_j)\,\mathrm{IW}(\Sigma_j; \nu, S),$$

with $q > 1$, which gives

$$\omega(\boldsymbol{\vartheta}_k) = \frac{1}{C_k}\prod_{1\le i<j\le k}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g}\prod_{j=1}^{k}\frac{N(\boldsymbol{\mu}_j; \mathbf{0}, gA_\Sigma)}{N(\boldsymbol{\mu}_j; \mathbf{0}, g\Sigma_j)}.$$

For the MOM-Beta in (2.1.10) $\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta}; q)\prod_{j=1}^{k}\prod_{f=1}^{p}\mathrm{Beta}(\theta_{jf}; ag, (1-a)g)$, hence

$$\omega(\boldsymbol{\vartheta}_k) = \frac{1}{C_k}\prod_{1\le i<j\le k}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j).$$

Our strategy is admittedly simple but has convenient advantages. After obtaining $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$ one need only compute a posterior average, which relative to the cost of $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$ is negligible. Furthermore, only posterior sampling under $\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ is required. As a potential caveat the posterior variance of $\omega(\boldsymbol{\vartheta}_k)$ has an effect on $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$, specifically when the local and non-local posteriors differ substantially this variance may be large.

However from Theorem 1 these posteriors differ mainly in overfitted mixtures ($k > k^*$), and only the numerator but not the denominator in $w(\boldsymbol{\vartheta}_k)$ may vanish (provided $\tilde{p}$ is positive over its domain, as is the case), hence in practice we found (4.1.1) to be quite stable as we discuss in Section 4.3).

In our examples estimates of $p(\mathcal{M}_k \mid \mathbf{y})$ were more precise than those of $\tilde{p}(\mathcal{M}_k \mid \mathbf{y})$, due to the former being closer to 0 or 1 (as expected from the posterior

concentration result in Theorem 1). Alternatively one can devise a sampler directly for the non-local $p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$, *e.g.* using slice sampling (Petralia et al., 2012), latent truncations (Rossell and Telesca, 2017) or collapsed Gibbs (Xie and Xu, 2017), but we do not pursue this as our main interest is model selection (see Section 4.2 for point estimates).

We now review computational approximations to estimate $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$. One option is to use trans-dimensional Markov chain Monte Carlo as in Richardson and Green (1997). Marin and Robert (2008) argue that this approach can require non-trivial calibration and that when k is small it may be preferable to explore each model separately. Related strategies are the dual importance sampling of (Lee and Robert, 2016) and the trans-dimensional collapsed Gibbs sampler by Xie and Xu (2017). Here we build on a refinement of an algorithm by Chib (1995) based on MCMC output. Neal (1999) showed that the algorithm fails when the sampler does not explore the $k!$ modes, hence the need for a correction (Berkhof et al., 2003; Marin and Robert, 2008). Specifically we use

$$\begin{aligned}\tilde{p}(\mathbf{y} \mid \mathcal{M}_k) &= \frac{p(\mathbf{y} \mid \hat{\boldsymbol{\vartheta}}_k, \mathcal{M}_k)\tilde{p}(\hat{\boldsymbol{\vartheta}}_k \mid \mathcal{M}_k)}{\tilde{p}(\hat{\boldsymbol{\vartheta}}_k \mid \mathbf{y}, \mathcal{M}_k)} \\ &= \frac{p(\mathbf{y} \mid \hat{\boldsymbol{\vartheta}}_k, \mathcal{M}_k)\tilde{p}(\hat{\boldsymbol{\vartheta}}_k \mid \mathcal{M}_k)}{\sum_{\psi \in \mathfrak{N}(k)} \tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \mathcal{M}_k)/(k!)}. \end{aligned} \tag{4.1.2}$$

For $\hat{\boldsymbol{\vartheta}}_k$ we use the posterior mode as recommended in Marin and Robert (2008). The numerator in (4.1.2) simply requires evaluating the likelihood and prior at $\hat{\boldsymbol{\vartheta}}_k$. To evaluate the denominator we note that under exchangeable $\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ the posterior is invariant to label-switching, thus

$$\tilde{p}(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) = \frac{1}{k!} \sum_{\psi \in \mathfrak{N}(k)} \tilde{p}(\psi(\boldsymbol{\vartheta}_k) \mid \mathbf{y}, \mathcal{M}_k), \tag{4.1.3}$$

where $\mathfrak{N}(k)$ is the set of $k!$ possible permutations of the set $\{1, ..., k\}$. Using a standard Rao-Blackwell argument as in Marin and Robert (2008) and defining the latent indicator z_i where $z_i = j$ if observation i is assigned to component j, we estimate (4.1.3) by

$$\frac{1}{Tk!} \sum_{\psi \in \mathfrak{N}(k)} \sum_{t=1}^{T} \tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \mathbf{z}^{(t)}, \mathcal{M}_k), \tag{4.1.4}$$

where $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_n^{(t)})$ are posterior samples from $\tilde{p}(\mathbf{z}, \boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$. The algorithm can be applied to any model subject to $\tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \mathbf{z}^{(t)}, \mathcal{M}_k)$ having closed-

form. $\tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \boldsymbol{z}^{(t)}, \mathcal{M}_k)$ is the complete posterior distribution under the cluster allocations, specifically

$$\tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \boldsymbol{z}^{(t)}, \mathcal{M}_k) = \prod_{j=1}^{k} N\left(\psi(\hat{\boldsymbol{\mu}}_j); \frac{gn_j^{(t)}\bar{\mathbf{y}}_j^{(t)}}{1+gn_j^{(t)}}, \frac{g}{1+gn_j^{(t)}}\Sigma_j^{(t)}\right) \text{IW}\left(\psi(\hat{\Sigma}_j); \nu + n_j^{(t)}, S_j^{(t)}\right)$$
$$\times \text{Dir}(\psi(\hat{\boldsymbol{\eta}}); q + n_1^{(t)}, ..., q + n_k^{(t)}),$$

$$\tilde{p}(\psi(\hat{\boldsymbol{\vartheta}}_k) \mid \mathbf{y}, \boldsymbol{z}^{(t)}, \mathcal{M}_k) = \prod_{j=1}^{k}\prod_{f=1}^{p} \text{Beta}\left(\psi(\hat{\theta}_{jf}); ag + \sum_{z_i^{(t)}=j} y_{if}, (1-a)g + \sum_{z_i^{(t)}=j}(L_{if} - y_{if})\right)$$
$$\times \text{Dir}(\psi(\hat{\boldsymbol{\eta}}); q + n_1^{(t)}, ..., q + n_k^{(t)}).$$

for the Normal and product Binomial mixtures, respectively. Algorithms 1-2 show in detail how to obtain $\hat{p}$ for Normal and product Binomial mixtures.

Algorithm 1: $p(\mathbf{y} \mid \mathcal{M}_k)$ for Normal mixtures under the MOM-IW prior.

Initialize $\boldsymbol{\vartheta}_k^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)}, \boldsymbol{\eta}^{(0)})$ with $\boldsymbol{\theta}_j^{(0)} = (\boldsymbol{\mu}_j^{(0)}, \Sigma_j^{(0)})$. **for** $t = 1, ..., T$ **do**

Draw $z_i^{(t)} = j$ with probability:

$$\frac{\eta_k^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}.$$

Let $n_j^{(t)} = \sum_{i=1}^{n} \text{I}(z_i^{(t)} = j)$ and $\bar{\mathbf{y}}_j^{(t)} = \frac{1}{n_j}\sum_{z_i^{(t)}=j} \mathbf{y}_i$ if $n_j^{(t)} > 0$, else $\bar{\mathbf{y}}_j^{(t)} = 0$. Draw

$$\boldsymbol{\eta}^{(t)} \sim \text{Dir}(q + n_1^{(t)}, ..., q + n_k^{(t)}).$$

Let $S_j = S^{-1} + \sum_{z_i=j}(\mathbf{y}_i - \bar{\mathbf{y}}_j^{(t-1)})(\mathbf{y}_i - \bar{\mathbf{y}}_j^{(t-1)})' + \sum_{j=1}^{k} \frac{n_j/g}{n_j + 1/g}\bar{\mathbf{y}}_j^{(t)}\bar{\mathbf{y}}_j^{'(t)}$. Draw

$$\Sigma_j^{(t)} \sim \text{IW}\left(\nu + n_j, S_j\right),$$

Draw

$$\boldsymbol{\mu}_j^{(t)} \sim N\left(\frac{gn_j^{(t)}\bar{\mathbf{y}}_j^{(t)}}{1+gn_j^{(t)}}, \frac{g}{1+gn_j^{(t)}}\Sigma_j^{(t)}\right),$$

end

Compute (4.1.1) and (4.1.2) where $\hat{\boldsymbol{\vartheta}}_k$ is the posterior mode under LPs.

Algorithm 2: $p(\mathbf{y} \mid \mathcal{M}_k)$ for product Binomial mixtures under the MOM-Beta prior.

Initialize $\boldsymbol{\vartheta}_k^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)}, \boldsymbol{\eta}^{(0)})$ where $\boldsymbol{\theta}_j^{(0)} = (\theta_{j1}^{(0)}, \ldots, \theta_{jp}^{(0)})$. **for** $t = 1, ..., T$ **do**

Draw $z_i^{(t)} = j$ with probability:

$$\frac{\eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}.$$

Draw

$$\boldsymbol{\eta}^{(t)} \sim \text{Dir}(q + n_1^{(t)}, ..., q + n_k^{(t)}).$$

where $n_j^{(t)} = \sum_{i=1}^{n} \text{I}(z_i^{(t)} = j)$. Draw

$$\theta_{jf}^{(t)} \sim \text{Beta}\left(ag + \sum_{z_i^{(t)}=j} y_{if}, (1-a)g + \sum_{z_i^{(t)}=j} (L_{if} - y_{if}) \right),$$

end

Compute (4.1.1) and (4.1.2) where $\hat{\boldsymbol{\vartheta}}_k$ is the posterior mode under LPs.

4.2 Posterior mode estimation

The EM algorithm provides a fast way to obtain posterior modes $\hat{\boldsymbol{\vartheta}}_k = \arg\max_{\boldsymbol{\vartheta}_k} p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ or cluster assigments $\hat{z}_i = \arg\max_{j \in \{1,...,K\}} p(z_i = j \mid \mathbf{y}, \hat{\boldsymbol{\vartheta}}_k, \mathcal{M}_k)$. We now briefly describe the algorithm. At iteration t the E-step computes

$$\bar{z}_{ij}^{(t)} = P(z_i = j \mid \mathbf{y}_i, \boldsymbol{\vartheta}_j^{(t-1)}) = \eta_j^{(t-1)} p(\mathbf{y}_i \mid \boldsymbol{\theta}_j^{(t-1)}) / \sum_{j=1}^{k} \eta_j^{(t-1)} p(\mathbf{y}_i \mid \boldsymbol{\theta}_j^{(t-1)})$$

and is trivial to implement. The M-step requires updating $\boldsymbol{\vartheta}_k^{(t)}$ in a manner that increases the expected complete log-posterior, which we denote by $\xi(\boldsymbol{\vartheta}_k)$, but under our prior $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ in general this cannot be done in closed-form. A key observation is that if $p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ leads to closed-form updates, the corresponding target $\xi^L(\boldsymbol{\vartheta}_k)$ only differs from $\xi(\boldsymbol{\vartheta}_k)$ by a term $d_\vartheta(\boldsymbol{\vartheta}_k)$, thus one may approximate $\xi(\boldsymbol{\vartheta}_k)$ via a first order Taylor expansion of $d_\vartheta(\boldsymbol{\vartheta}_k)$.

These approximate updates need not lead to an increase in $\xi(\boldsymbol{\vartheta}_k)$ (although they typically do since $d_\vartheta(\boldsymbol{\vartheta}_k)$ has a mild influence for moderately large n), whenever this happens we use gradient algorithm updates. We now describe the derivation of the Algorithm 3 for Normal mixtures under MOM-IW priors and Algorithm 4 for product Binomial mixtures under MOM-Beta priors.

Algorithm 3: EM under MOM-IW-Dir priors.

Set $t = 1$. **while** $\zeta > \epsilon^*$ *and* $t < T$ **do**

for $t \geq 1$ *and* $j = 1, ..., k$ **do**

E-step. Let $\bar{z}_{ij}^{(t)} = \dfrac{\eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^k \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}$ and $n_j^{(t)} = \sum_{i=1}^n \bar{z}_{ij}^{(t)}$.

M-step. Let $\bar{\mathbf{y}}_j^{(t)} = \sum_{i=1}^n \bar{z}_{ij}^{(t)} \mathbf{y}_i / n_j^{(t)}$.

Update

$$\boldsymbol{\mu}_j^{(t)} = \left(\left(\Sigma_j^{-1}\right)^{(t-1)} n_j^{(t)} + A^{-1}_{\Sigma^{(t-1)}} \left(\frac{1}{g} + \sum_{i \neq j} \frac{2}{d_{ij}} \right) \right)^{-1} \times \left(\Sigma^{-1(t-1)} n_j^{(t)} \bar{\mathbf{y}}_j^{(t)} + A^{-1}_{\Sigma^{(t-1)}} \left(\sum_{i \neq j} \frac{\boldsymbol{\mu}_j^{(t-1)} - (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})}{d_{ij}} \right) \right),$$

Update

$$(\nu - p + n_j^{(t)}) \Sigma_j^{(t)} = S^{-1} + \frac{\boldsymbol{\mu}_j^{(t)} (\boldsymbol{\mu}_j^{(t)})'}{kg} + \sum_{i=1}^n \bar{z}_{ij}^{(t)} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})' - \frac{1}{k} \sum_{i \neq j} \frac{2(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})'}{d_{ij}}.$$

Update $\eta_j^{(t)} = \dfrac{n_j^{(t)} + q - 1}{n + k(q-1)}$.

end

Compute $\zeta = |\xi(\boldsymbol{\vartheta}_k^{(t)}) - \xi(\boldsymbol{\vartheta}_k^{(t-1)})|$ and set $t = t + 1$.

end

Algorithm 4: EM under MOM-Beta priors.

Set $t = 1$. **while** $\zeta > \epsilon^*$ *and* $t < T$ **do**

for $t \geq 1$ *and* $j = 1, ..., k$ **do**

E-step. Let $\bar{z}_{ij}^{(t)} = \dfrac{\eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}$.

M-step. Update

$$\boldsymbol{\theta}_j^{(t)} = \frac{ag + \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} \mathbf{y}_i + \ell_1(\boldsymbol{\theta}_j^{(t)})}{(1-a)g + \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} (L_{if} - \mathbf{y}_i) + 2\ell_2(\boldsymbol{\theta}_j^{(t)})},$$

$\ell_1(\boldsymbol{\theta}_j^{(t)}) = \dfrac{\boldsymbol{\theta}_j^{(t-1)} - (\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})}{(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})'(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})}$,

$\ell_2(\boldsymbol{\theta}_j^{(t)}) = \left[(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})'(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})\right]^{-1}$. Update

$\eta_j^{(t)} = \dfrac{n_j^{(t)} + q - 1}{n + k(q-1)}$.

end

Compute $\zeta = |\xi(\boldsymbol{\vartheta}_k^{(t)}) - \xi(\boldsymbol{\vartheta}_k^{(t-1)})|$ and set $t = t + 1$.

end

Algorithm 5: Gradient Ascend algorithm.

Initialization $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*$, $\bar{k} = \sqrt{\dfrac{\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}^{(t-1)}\|}{\nabla\xi(\boldsymbol{\zeta}^{(t-1)})}}$ and $h = 0$;

while $(\xi(\boldsymbol{\zeta}^{(t-1)}) > \xi(\boldsymbol{\zeta}^*))$ **do**

  $\boldsymbol{\zeta}^* = \boldsymbol{\zeta}^{(t-1)} + \dfrac{\bar{k}}{2^h} \nabla\xi(\boldsymbol{\zeta}^{(t-1)})$;

  $h = h + 1$

end

$\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}^*$

4.2.1 Derivation of the EM algorithm for Normal mixtures under MOM-IW-Dir priors

The complete-data posterior can be written as follows

$$p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathbf{z}, \mathcal{M}_k) = \frac{1}{C_k} \prod_{1 \le i < j \le k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \tag{4.2.1}$$
$$\times \prod_{j=1}^{k} \prod_{i=1}^{n} (\eta_j N(\mathbf{y}; \boldsymbol{\mu}_j, \Sigma_j))^{z_{ij}} N\left(\boldsymbol{\mu}_j; \mathbf{0}, g A_\Sigma\right) \text{Wishart}(\Sigma_j^{-1}; \nu, S) \text{Dir}(\boldsymbol{\eta}; q).$$

The E-step at iteration t requires the expectation of $\log p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathbf{z}, \mathcal{M}_k)$ with respect to $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\vartheta}_k^{(t-1)}, \mathcal{M}_k)$, where $\boldsymbol{\vartheta}_k^{(t-1)} = (\boldsymbol{\eta}^{(t-1)}, \boldsymbol{\mu}_1^{(t-1)}, ..., \boldsymbol{\mu}_k^{(t-1)}, \Sigma_1^{(t-1)}, ..., \Sigma_k^{(t-1)})$ are the parameter values at iteration $t-1$. Let

$$\bar{z}_{ij}^{(t)} = p(z_i = j \mid \mathbf{y}_i, \boldsymbol{\vartheta}_k^{(t-1)}) = \frac{\eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} N(\mathbf{y}_i; \boldsymbol{\mu}_j^{(t-1)}, \Sigma_j^{(t-1)})}, \tag{4.2.2}$$

then the M-step seeks $\boldsymbol{\vartheta}_k^{(t)}$ maximising

$$\log(p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \bar{z}_{ij}, \mathcal{M}_k)) = \sum_{j=1}^{k} n_j \log(\eta_j) + \sum_{j=1}^{k} \sum_{i=1}^{n} \bar{z}_{ij} \log(N(\boldsymbol{y}_i; \boldsymbol{\mu_j}, \Sigma_j)) \tag{4.2.3}$$
$$+ \sum_{j=1}^{k} \log(N\left(\boldsymbol{\mu}_j; \mathbf{0}, g A_\Sigma\right)) + \sum_{j=1}^{k} \log(\text{Wishart}(\Sigma_j^{-1}; \nu, S))$$
$$+ \sum_{1 \le i < j \le k} \log((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)) + \log(\text{Dir}(\boldsymbol{\eta}; q))$$

where $n_j^{(t)} = \sum_{i=1}^{n} \bar{z}_{ij}^{(t)}$. We successively update $\boldsymbol{\eta}^{(t)}$, $\boldsymbol{\mu}_1^{(t)}$,...,$\boldsymbol{\mu}_k^{(t)}$ and $\Sigma_1^{(t)}, ..., \Sigma_k^{(t)}$ in a fashion that guarantees that (4.2.3) increases at each step. The update $\eta_j^{(t)}$ is

$$\eta_j^{(t)} = \frac{n_j^{(t)} + q - 1}{n + k(q-1)}, \tag{4.2.4}$$

which maximizes (4.2.3) with respect to $\boldsymbol{\eta}$ conditional on the current $\boldsymbol{\mu}_1^{(t-1)}$,...,$\boldsymbol{\mu}_k^{(t-1)}$ and $\Sigma_1^{(t-1)}, ..., \Sigma_k^{(t-1)}$. To update $\boldsymbol{\mu}_j^{(t)}$ we seek to maximize

$$\xi(\boldsymbol{\mu}_j^{(t)}) = \sum_{i \neq j} \log(\boldsymbol{C}_{ij}^{(t)'}(\Sigma_j^{(t-1)})^{-1}\boldsymbol{C}_{ij}^{(t)}) \\ - \frac{1}{2g}\boldsymbol{\mu}_j^{'(t)}A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{\mu}_j^{(t)} - \frac{1}{2}\sum_{i=1}^{n}\bar{z}_{ij}^{(t)}(\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)})'(\Sigma^{-1})^{(t-1)}(\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)}),$$

where $\boldsymbol{C}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. The first derivative of $\xi(\boldsymbol{\mu}_j^{(t)})$ is

$$\nabla\xi(\boldsymbol{\mu}_j^{(t)}) = -2\sum_{i \neq j}\frac{A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{C}_{ij}^{(t)}}{\boldsymbol{C}_{ij}^{(t)'}A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{C}_{ij}^{(t)}} - \frac{1}{g}(A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{\mu}_j^{(t)}) - \sum_{i=1}^{n}\bar{z}_{ij}^{(t)}(A_{\Sigma_j^{(t-1)}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(t)})).$$

Because an analytic solution of $\nabla\xi(\boldsymbol{\mu}_j^{(t)}) = \mathbf{0}$ in terms of $\boldsymbol{\mu}_j^{(t)}$ is not feasible we resort to a first order Taylor approximation for $-2\sum_{i\neq j}(A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{C}_{ij}^{(t)})/(\boldsymbol{C}_{ij}^{(t)'}A_{\Sigma^{(t-1)}}^{-1}\boldsymbol{C}_{ij}^{(t)})$ around $\boldsymbol{\mu}_j^{(t-1)}$. Finding the maximum of this Taylor approximation gives the candidate update

$$\boldsymbol{\mu}_j^* = \left(\Sigma_j^{-1(t-1)}n_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1}\left(\frac{1}{g} + \sum_{j\neq k}\frac{2}{d_{ij}^{(t-1)}}\right)\right)^{-1} \\ \times\left(\Sigma^{-1(t-1)}n_j^{(t)}\bar{\mathbf{y}}_j^{(t)} + A_{\Sigma^{(t-1)}}^{-1}\left(\sum_{i\neq j}\frac{\boldsymbol{\mu}_j^{(t-1)} - (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})}{d_{ij}^{(t-1)}}\right)\right), \tag{4.2.5}$$

where $d_{ij}^{(t-1)} = (\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})'A_{\Sigma^{(t-1)}}^{-1}(\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})$. If $\xi(\boldsymbol{\mu}_j^*) > \xi(\boldsymbol{\mu}_j^{(t-1)})$ we set $\boldsymbol{\mu}_j^{(t)} = \boldsymbol{\mu}_j^*$, else take the gradient step in Algorithm 5.

Finally we describe updating Σ_j for $j = 1, \ldots, k$. Redefine $\xi(\Sigma_j)$ to now be (4.2.3) viewed as a function Σ_j. Due to the terms $\sum_{i\neq j}\log(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)})'A_{\Sigma^{(t)}}^{-1}(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)})$ and $-\frac{1}{2}\log(|A_{\Sigma^{(t)}}^{-1}|)$ an analytic solution of $\nabla\xi(\Sigma_j) = \mathbf{0}$ is not available, hence we use the Taylor expansion around $\Sigma_j^{(t-1)}$

$$\sum_{i\neq j}\log(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)})'A_{\Sigma^{(t)}}^{-1}(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)}) - \frac{1}{2}\log(|A_{\Sigma^{(t)}}^{-1}|) \approx \\ \sum_{i\neq j}\frac{(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)})'A_{\Sigma^{(t)}}^{-1}(\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)})}{(\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})'A_{\Sigma^{(t-1)}}^{-1}(\boldsymbol{\mu}_i^{(t-1)} - \boldsymbol{\mu}_j^{(t-1)})} - \frac{1}{2}\log(|\Sigma_j^{(t)}|).$$

Note that when a common $\Sigma_1 = \ldots = \Sigma_k$ is assumed then $A_{\Sigma^{(t)}} = \Sigma^{(t)}$ we

only need a Taylor expansion of first term. Summarising, the candidate update is

$$(\nu - p + n_j^{(t)})\Sigma_j^* = S^{-1} + \frac{\boldsymbol{\mu}_j^{(t)}(\boldsymbol{\mu}_j^{(t)})'}{kg} + \sum_{i=1}^{n} \bar{z}_{ij}^{(t)}(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})' - \frac{1}{k}\sum_{i\neq j} \frac{2(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_k^{(t)})'}{d_{ij}^{(t-1)}}.$$

If $\xi(\Sigma_j^*) > \xi(\Sigma_j^{(t-1)})$ we set $\Sigma_j^{(t)} = \Sigma_j^*$, else take a gradient step (Algorithm 5) with a small enough step size to ensure that $\Sigma_j^{(t)}$ remains positive-definite.

4.2.2 Derivation of the EM algorithm for product Binomial mixtures under MOM-Beta-Dir priors

The complete-data posterior can be written as follows

$$p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \boldsymbol{z}, \mathcal{M}_k) = \prod_{j=1}^{k}\prod_{i=1}^{n}(\eta_j \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}))^{z_{ij}} \frac{1}{C_k} \prod_{1\leq i<j\leq k} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \times \prod_{f=1}^{p} \text{Beta}(\theta_{jf}; ag, (1-a)g)\text{Dir}(\boldsymbol{\eta}; q). \tag{4.2.6}$$

For the t-th iteration of the E-step, we compute the expectation of the latent cluster allocations $\bar{z}_{ij}^{(t)} = p(z_i = j \mid \mathbf{y}_i, \boldsymbol{\vartheta}_j^{(t-1)})$ given $\boldsymbol{\eta}^{(t-1)}$ and $\boldsymbol{\theta}_1^{(t-1)}$,...,$\boldsymbol{\theta}_k^{(t-1)}$ using

$$\bar{z}_{ij}^{(t)} = \frac{\eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}{\sum_{j=1}^{k} \eta_j^{(t-1)} \prod_{f=1}^{p} \text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})}. \tag{4.2.7}$$

For the t-th iteration of the M-step, we find the maximizers $\boldsymbol{\eta}^{(t)}$ and $\boldsymbol{\theta}_1^{(t)}$,...,$\boldsymbol{\theta}_k^{(t)}$ given $\bar{z}_{ij}^{(t)}$ of the following function

$$\begin{aligned}\log(p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \bar{z}_{ij}, \mathcal{M}_k)) = &\sum_{j=1}^{k} n_j \log(\eta_j) + \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{f=1}^{p} \bar{z}_{ij} \log(\text{Bin}(y_{if}; L_{if}, \theta_{jf}^{(t-1)})) \\ &+ \sum_{j=1}^{k}\sum_{f=1}^{p} \log(\text{Beta}(\theta_{jf}; ag, (1-a)g)) \\ &+ \sum_{1\leq i<j\leq k} \log(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) + \log(\text{Dir}(\boldsymbol{\eta}; q)) + \text{Constant},\end{aligned} \tag{4.2.8}$$

with $n_j^{(t)} = \sum_{i=1}^{n} \bar{z}_{ij}^{(t)}$. We update $\boldsymbol{\eta}^{(t)}$ and $\boldsymbol{\theta}_1^{(t)}$,...,$\boldsymbol{\theta}_k^{(t)}$ in a fashion that guarantees that (4.2.8) increases at each step. The M-step for $\eta_j^{(t)}$ is computed by using

$$\eta_j^{(t)} = \frac{n_j^{(t)} + q - 1}{n + k(q-1)}, \tag{4.2.9}$$

which maximizes (4.2.8) with respect to $\boldsymbol{\eta}$ conditional on the current $\boldsymbol{\theta}_1^{(t-1)}$,...,$\boldsymbol{\theta}_k^{(t-1)}$. For $\boldsymbol{\theta}_j^{(t)}$ let

$$\begin{aligned}\xi(\boldsymbol{\theta}_j^{(t)}) = &\sum_{i \neq j} \log(\boldsymbol{C}_{ij}^{(t)'} \boldsymbol{C}_{ij}^{(t)}) + \sum_{f=1}^{p} \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} y_{if} \log(\theta_{jf}^{(t)}) \\ &+ \sum_{i=1}^{n} \sum_{f=1}^{p} \bar{z}_{ij}^{(t)} (L_{if} - y_{if}) \log(1 - \theta_{jf}^{(t)}) + ((1-a)g - 1) \sum_{f=1}^{p} \log(1 - \theta_{jf}^{(t)}) \\ &+ (ag - 1) \sum_{f=1}^{p} \log(\theta_{jf}^{(t)}),\end{aligned}$$

be the corresponding target where $\boldsymbol{C}_{ij} = (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)$. The first derivative of $\xi(\boldsymbol{\theta}_j^{(t)})$ is

$$\begin{aligned}\nabla \xi(\boldsymbol{\theta}_j^{(t)}) = &-2 \sum_{i \neq j} \frac{\boldsymbol{C}_{ij}^{(t)}}{\boldsymbol{C}_{ij}^{(t)'} \boldsymbol{C}_{ij}^{(t)}} + \sum_{i=1}^{n} \sum_{f=1}^{p} \bar{z}_{ij}^{(t)} y_{if} / \theta_{jf}^{(t)} \\ &- \sum_{i=1}^{n} \sum_{f=1}^{p} \bar{z}_{ij}^{(t)} (L_{if} - y_{if})) / (1 - \theta_{jf}^{(t)}) - \sum_{f=1}^{p} ((1-a)g - 1) / (1 - \theta_{jf}^{(t)}) \\ &+ \sum_{f=1}^{p} (ag - 1) / \theta_{jf}^{(t)}.\end{aligned}$$

An analytic solution of $\nabla \xi(\boldsymbol{\theta}_j^{(t)}) = \mathbf{0}$ in terms of $\boldsymbol{\theta}_j^{(t)}$ is not feasible. Hence we resort to a first order Taylor approximation for $-2 \sum_{i \neq j} (\boldsymbol{C}_{ij}^{(t)}) / (\boldsymbol{C}_{ij}^{(t)'} \boldsymbol{C}_{ij}^{(t)})$ around $\boldsymbol{\theta}_j^{(t)}$ and we now compute the M-step for $\boldsymbol{\theta}_j^*$ given by

$$\boldsymbol{\theta}_j^* = \frac{ag + \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} \mathbf{y}_i + \ell_1(\boldsymbol{\theta}_j^{(t)})}{(1-a)g + \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} (L_{if} - \mathbf{y}_i) + 2\ell_2(\boldsymbol{\theta}_j^{(t)})},$$

where

$$\ell_1(\boldsymbol{\theta}_j^{(t)}) = \frac{\boldsymbol{\theta}_j^{(t-1)} - (\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})}{(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})' (\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})},$$

and

$$\ell_2(\boldsymbol{\theta}_j^{(t)}) = \left[(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})'(\boldsymbol{\theta}_i^{(t-1)} - \boldsymbol{\theta}_j^{(t-1)})\right]^{-1}.$$

If $\xi(\boldsymbol{\theta}_j^*) > \xi(\boldsymbol{\theta}_j^{(t-1)})$ set $\boldsymbol{\theta}_j^{(t)} = \boldsymbol{\theta}_j^*$, else take the gradient step in Algorithm 5.

Algorithms 3 and 4 detail the steps for Normal and product Binomial mixtures (extensions to other models follow similar lines), for simplicity outlining the approximate updates given by the gradient Algorithm 5.

The proposed algorithm in this chapter for finding maximizers uses iteratively EM steps with gradient algorithm updates. The algorithm has some connections with MM algorithms (Lange et al., 2000; Hunter and Lange, 2004) which use convexity assumptions for finding maximizers and where the missing data structure is not necessary as in the EM algorithm. The algorithms in Lange et al. (2000) are based on a majorizing or minorizing function that serves as a surrogate for the objective function. A combination of the EM and MM algorithms is given in (Gormley and Murphy, 2008) where an Expectation Minorization Maximization (EMM) algorithm is proposed for model fitting in a mixture of experts model for rank data.

4.3 Precision of MCMC-based estimates relative to local priors

We compared empirically the precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ vs. the local prior-based $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$ for univariate and bivariate Normal mixtures with $k = 2, 3$ components (if $k = 1$ then $p(\mathbf{y} \mid \mathcal{M}_k) = p^L(\mathbf{y} \mid \mathcal{M}_k)$ has closed form). To inspect whether the precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ suffers under overfitted mixtures we simulated a single data set of $n = 500$ observations from a $k^* = 1$ component mixture and computed 100 times both $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ and $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$. Figures 4.1 and 4.2 show the results for a univariate and bivariate outcome respectively. The precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ was comparable to that of $\tilde{p}(\mathbf{y} \mid \mathcal{M}_k)$, in fact in some situations the former was more precise (this is due to $\text{Var}(\log \hat{p}) = \text{Var}(\log \tilde{p}) + \text{Var}(\log \hat{\omega}) + 2\text{cov}(\log \tilde{p}, \log \hat{\omega})$ where the latter covariance may be negative). More importantly, posterior model probabilities $\hat{p}(\mathcal{M}_k \mid \mathbf{y})$ (middle panels) were more precise than $\tilde{p}(\mathcal{M}_k \mid \mathbf{y})$, as in our experience tends to be the case due to $p(\mathcal{M}_k \mid \mathbf{y})$ having a higher concentration around 0 or 1 (Theorem 1). The lower panels show that as k grows larger than k^* the precision in $\hat{w}$ tends to degrade, however as mentioned this is compensated by the fact that $p(\mathcal{M}_k \mid \mathbf{y})$ is small for large k (middle panels), thus it does not appear to be a practical concern.

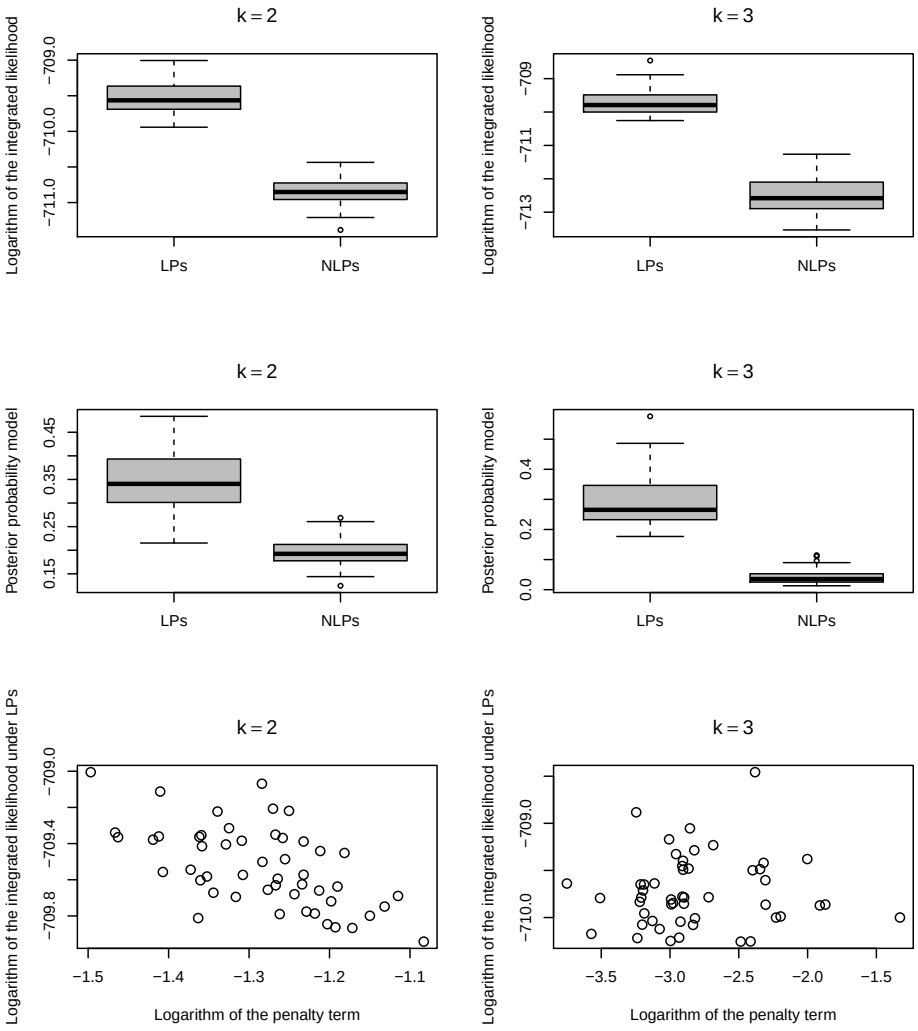


Figure 4.1: Precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ in 100 univariate simulations, $k^* = 1$. Top: $\log \hat{p}(\mathbf{y} \mid \mathcal{M}_k)$. Middle: $\hat{p}(\mathcal{M}_k \mid \mathbf{y})$. Bottom: $\log \hat{p}^L(\mathbf{y} \mid \mathcal{M}_k)$ vs. $\log \hat{E}(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y})$

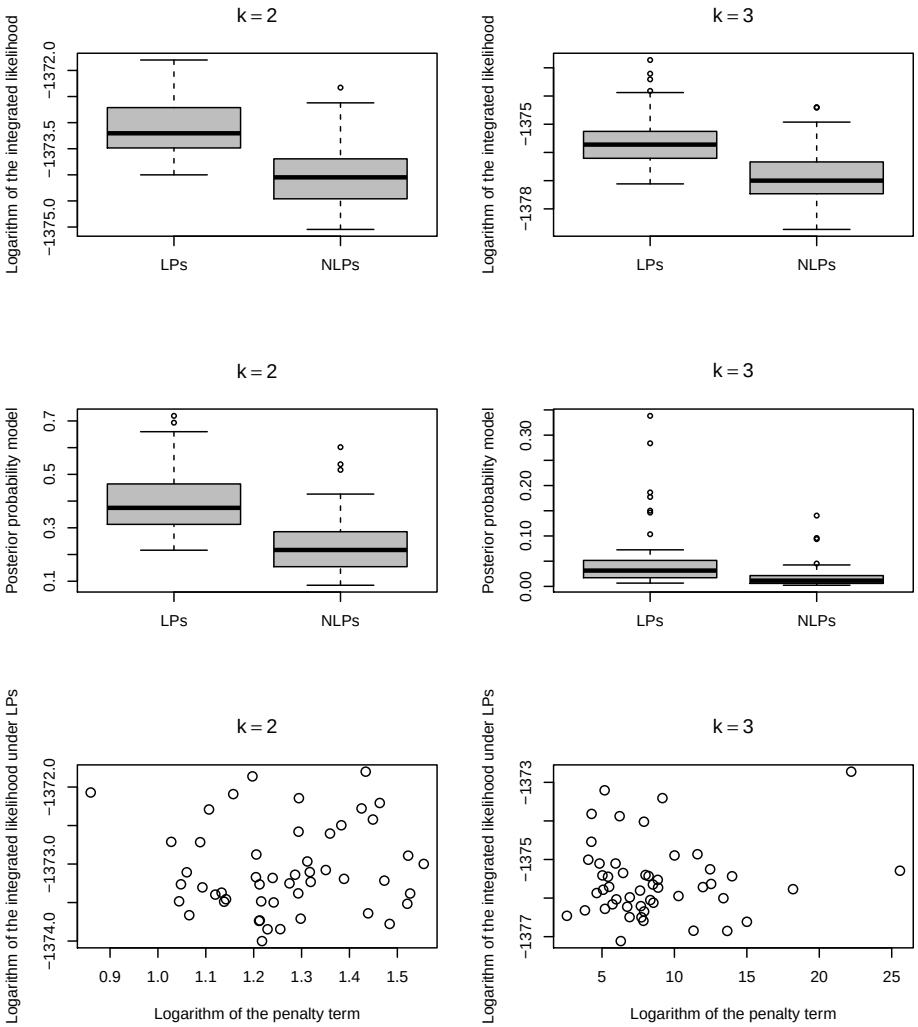


Figure 4.2: Precision of $\hat{p}(\mathbf{y} \mid \mathcal{M}_k)$ in 100 bivariate simulations, $k^* = 1$. Top: $\log \hat{p}(\mathbf{y} \mid \mathcal{M}_k)$. Middle: $\hat{p}(\mathcal{M}_k \mid \mathbf{y})$. Bottom: $\log \hat{p}^L(\mathbf{y} \mid \mathcal{M}_k)$ vs. $\log \hat{E}(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y})$

We now illustrate the performance of the EM algorithm for Normal mixture models under MOM-IW priors. Figures 4.3 and 4.4 show EM estimates for 300 data sets using Algorithm 3 from univariate and bivariate Normal mixtures. In Figure 4.3 for $k = 2$ the penalty and data variability seem to be negatively correlated (bottom panel) as we would expect considering the definition (2.1.4) for MOM-IW priors. A similar situation is illustrated in Figure 4.4 for $k = 2$ where we also see that the posterior modes recover the true parameter values (top panels).

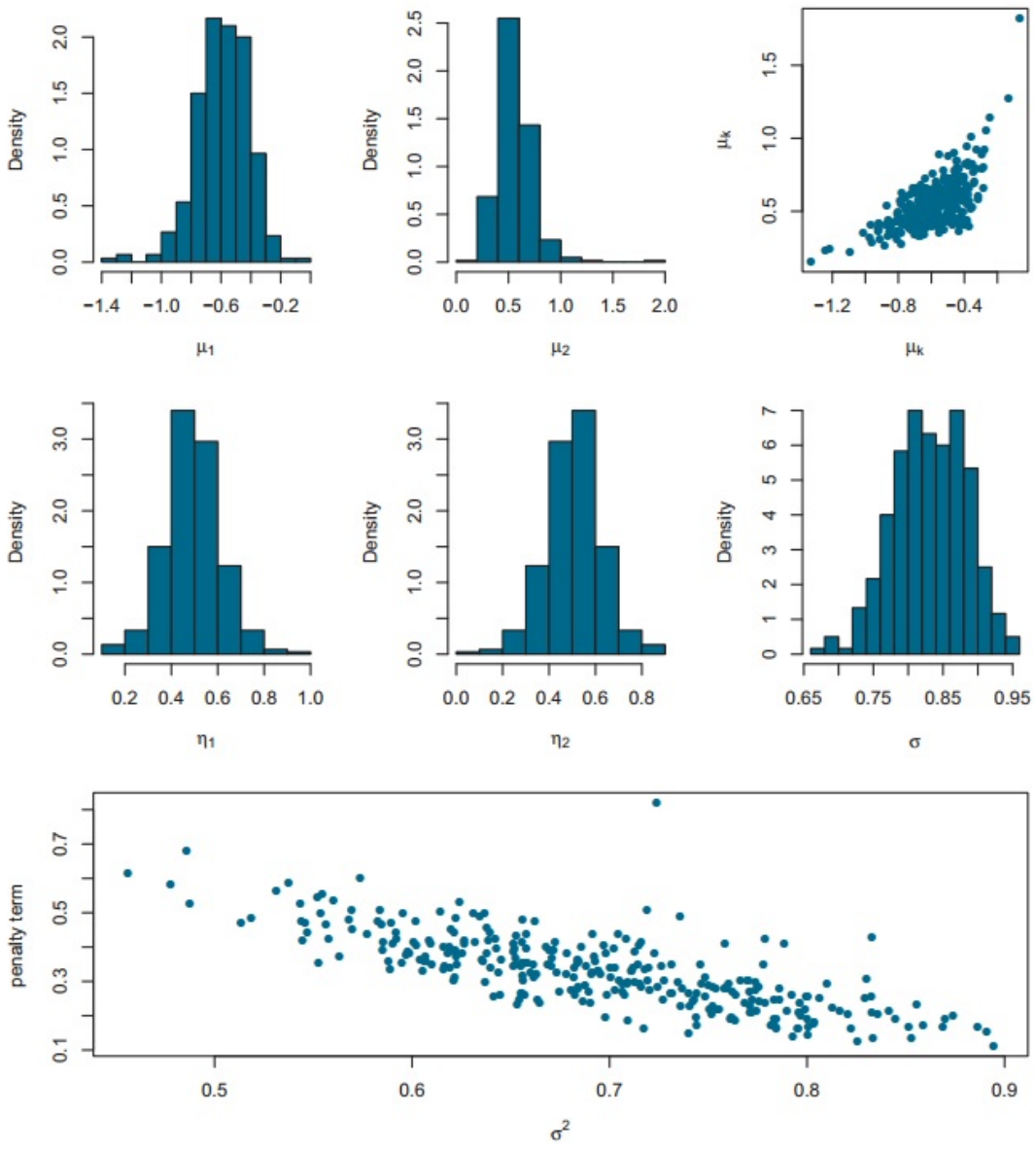


Figure 4.3: EM estimates for 300 data sets of $n = 500$ from a univariate Normal where $\mu = 0$ and $\sigma^2 = 1$.

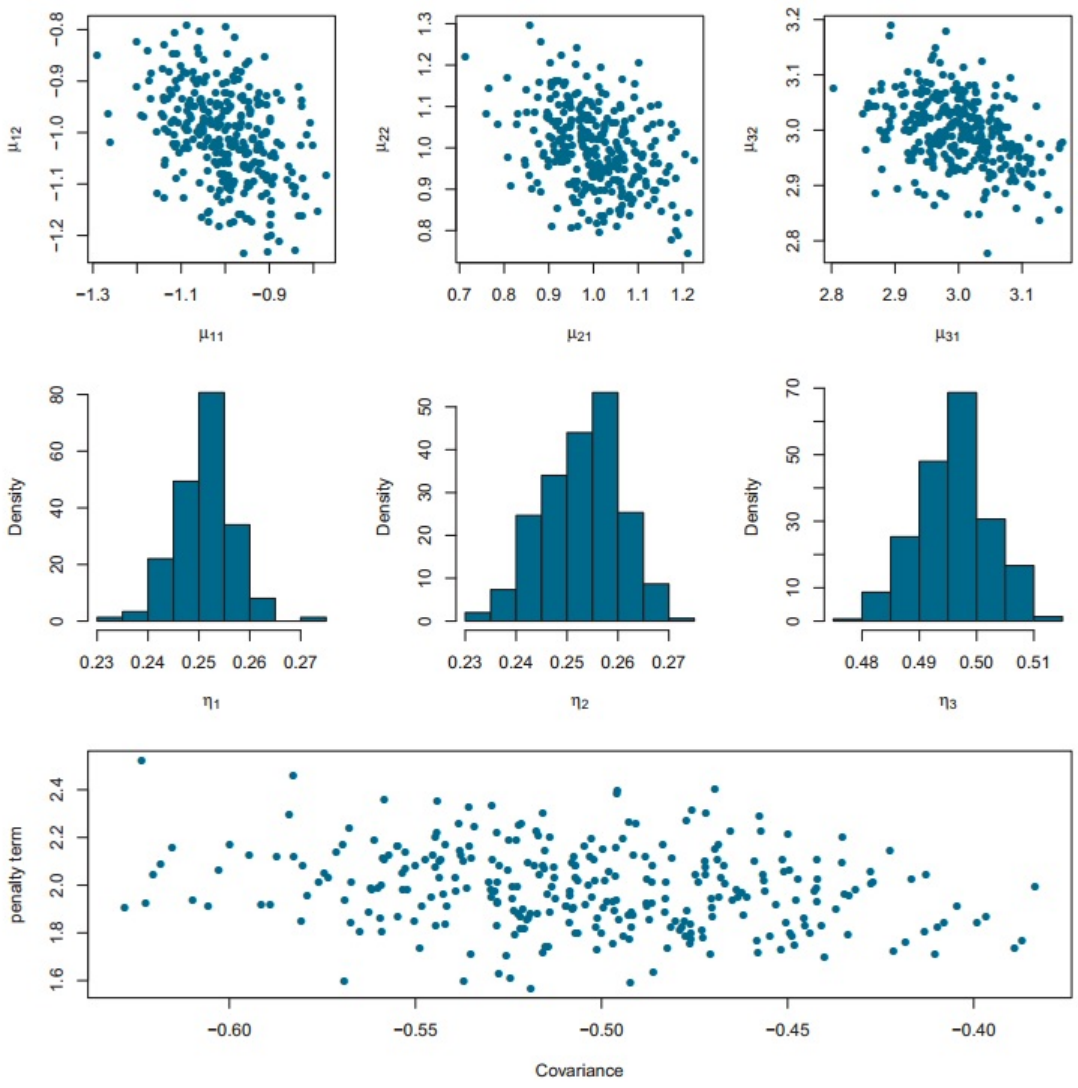


Figure 4.4: EM estimates for 300 data sets of $n = 500$ from a bivariate three component Normal mixture where $\boldsymbol{\mu}_1 = (-1,-1)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (3,3)'$, $\boldsymbol{\eta}$=(0.25,0.25,0.5), $\sigma^2_{11} = \sigma^2_{22} = 1$ and $\sigma^2_{12} = \sigma^2_{21} = -0.5$.

Chapter 5

Simulation studies

In this chapter, we compare our MOM-IW and MOM-Beta priors with default parameters to their local counterparts Normal-IW and Beta (respectively) with parameter g^L set to match the 95% percentile for the separation parameter κ (Chapter 3). Throughout, we set uniform prior model probabilities $P(\mathcal{M}_1) = \ldots = P(\mathcal{M}_k) = 1/k$. We estimate the integrated likelihoods using Algorithm 1 based on 5,000 MCMC draws after a 2,500 burn-in period. We also consider the BIC, AIC, sBIC as it is currently implemented by Drton and Plummer (2017) where $\Sigma_i \neq \Sigma_j$. Section 5.1 presents a simulation study for univariate and bivariate Normal mixtures. In Section 5.2 we explore model misspecification by simulating data from a T mixture and the two-piece iskew-T mixture proposed in Rossell and Steel (2017). Section 5.3 reproduces a Binomial mixture example by Drton and Plummer (2017) to illustrate the sBIC. In Section 5.4, we perform a sensitivity analysis for the hyperparameters of MOM-IW and MOM-Beta priors. Finally, Section 5.5 shows a simulation experiment for product Binomial mixtures which illustrates the usage of diagnostics for multiple EM and MCMC runs.

5.1 Normal mixture

Here, we consider a simulation study where the goal is to choose one amongst the three competing models

$$\begin{aligned}
&\mathcal{M}_1 : N(\mathbf{y}_i; \boldsymbol{\mu}, \Sigma), \\
&\mathcal{M}_2 : \eta_1 N(\mathbf{y}_i; \boldsymbol{\mu}_1, \Sigma) + (1-\eta_1) N(\mathbf{y}_i; \boldsymbol{\mu}_2, \Sigma) \\
&\mathcal{M}_3 : \eta_1 N(\mathbf{y}_i; \boldsymbol{\mu}_1, \Sigma) + \eta_2 N(\mathbf{y}_i; \boldsymbol{\mu}_2, \Sigma) + (1-\eta_1-\eta_2) N(\mathbf{y}_i; \boldsymbol{\mu}_3, \Sigma),
\end{aligned}$$

Table 5.1: Cases for the simulation study data-generating truth. $\Sigma = 1$ in Cases 1-4, $\sigma^2_{11} = \sigma^2_{22} = 1$ and $\sigma^2_{12} = \sigma^2_{21} = -0.5$ in Cases 5-8.

Case 1	k^*=1	$\boldsymbol{\mu}_1$=0	
Case 2	k^*=2	$\boldsymbol{\mu}_1$=-1, $\boldsymbol{\mu}_2$=1	$\boldsymbol{\eta} = (0.5, 0.5)$
Case 3	k^*=2	$\boldsymbol{\mu}_1 = -2, \boldsymbol{\mu}_2 = 2$	$\boldsymbol{\eta} = (0.5, 0.5)$
Case 4	k^*=3	$\boldsymbol{\mu}_1 = -1, \boldsymbol{\mu}_2 = 1, \boldsymbol{\mu}_3 = 4$	$\boldsymbol{\eta}$ =(0.45,0.45,0.1)
Case 5	k^*=1	$\boldsymbol{\mu} = (0, 0)'$	
Case 6	k^*=2	$\boldsymbol{\mu}_1 = (-0.4, -0.6)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$	$\boldsymbol{\eta} = (0.5, 0.5)$
Case 7	k^*=2	$\boldsymbol{\mu}_1 = (-0.65, -0.85)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$	$\boldsymbol{\eta}$=(0.35,0.35,0.3)
Case 8	k^*=3	$\boldsymbol{\mu}_1 = (-0.65, -0.85)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (3, 3)'$	$\boldsymbol{\eta}$=(0.35,0.35,0.3)

We simulated 100 datasets under each of the 8 data-generating truths with Normal components depicted in Figure 5.1 and Table 5.1 for univariate (Cases 1-4) and bivariate outcomes (Cases 5-8). Case 1 corresponds to $k^* = 1$ components, Cases 2-3 to $k^* = 2$ moderately and strongly-separated components respectively, and Case 4 to $k^* = 3$ with two strongly overlapping components and a third component with smaller weight. Cases 5-8 are analogous for the bivariate outcome.

Figures 5.2-5.3 show the average posterior probability assigned to the data-generating model $P(\mathcal{M}_{k^*} \mid \mathbf{y})$. To compare BIC, AIC model selection criteria and LPs, NLPs, Figures 5.4-5.5 report the proportion of correct model selections, *i.e.* the proportion of simulated datasets in which $\hat{k} = k^*$, where $\hat{k}$ is the selected number of components by any given method (for Bayesian methods $\hat{k} = \arg\max_k p(\mathcal{M}_k \mid \mathbf{y})$).

Overall a similar behavior is observed in the univariate and bivariate cases. The BIC adequately favoured sparse solutions (Cases 1,3,5,7) but showed an important lack of sensitivity to detect some truly present components (Cases 2,4,6,8). AIC was suboptimal in almost all scenarios.

As seen in Figures 5.2-5.3, the Normal-IW led to substantially less posterior concentration of $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ than our MOM-IW in all cases except the non-sparse Cases 4 and 8, where results were practically indistinguishable. As predicted by theory, the Normal-IW put too much posterior mass on overfitted models. Interestingly, Cases 2 and 6 illustrate that additionally to enforcing parsimony MOM-IW can sometimes also increase sensitivity to detect moderately-separated components. This is due to assigning larger prior $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ consistent with that degree of separation. Figures 5.13-5.18 show similar results, there $P(\kappa < 4 \mid \mathcal{M}_k) = 0.05$ led to slightly better parsimony than $P(\kappa < 4 \mid \mathcal{M}_k) = 0.10$.

where independence is assumed across $i = 1, \ldots, n$.

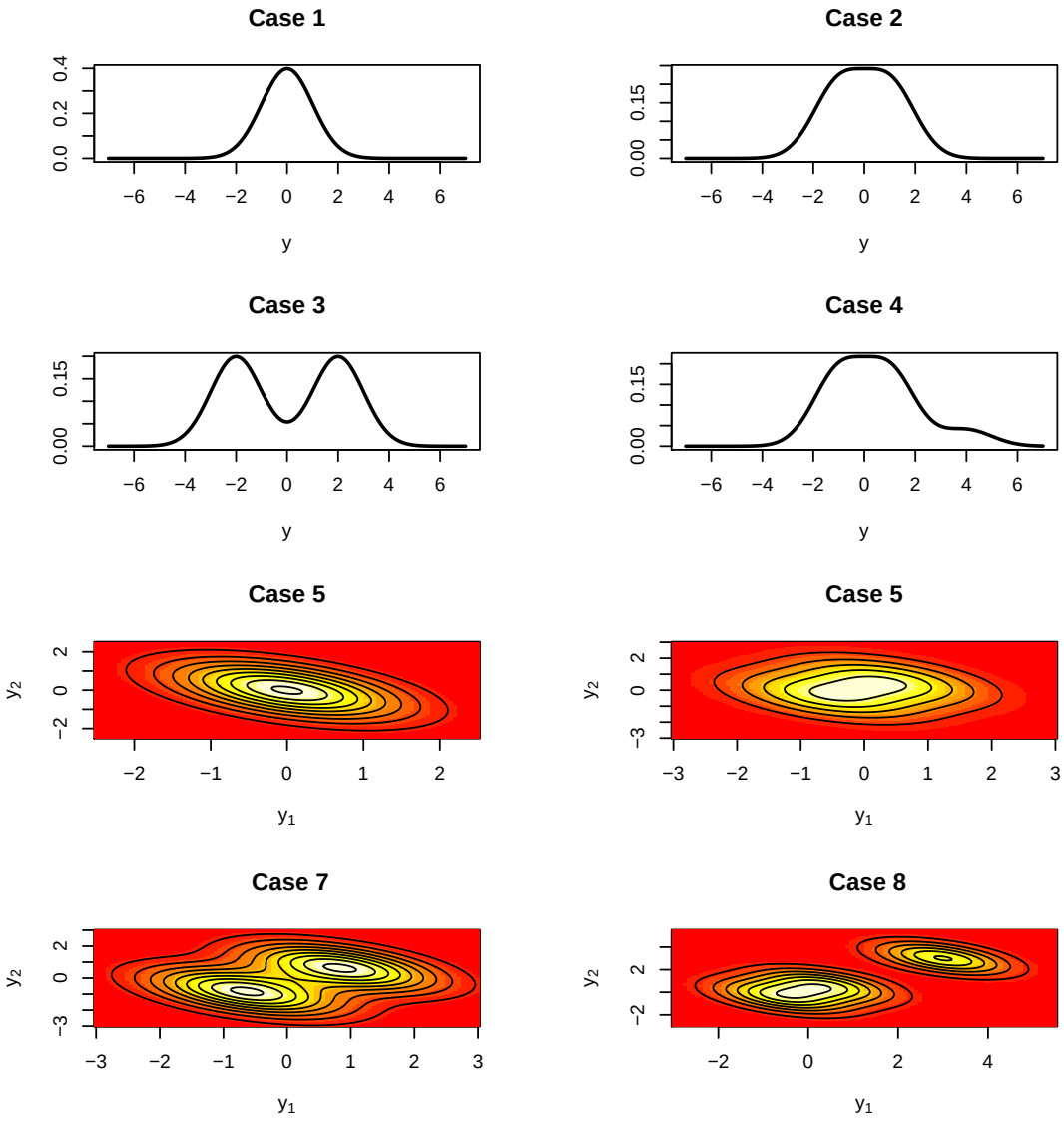
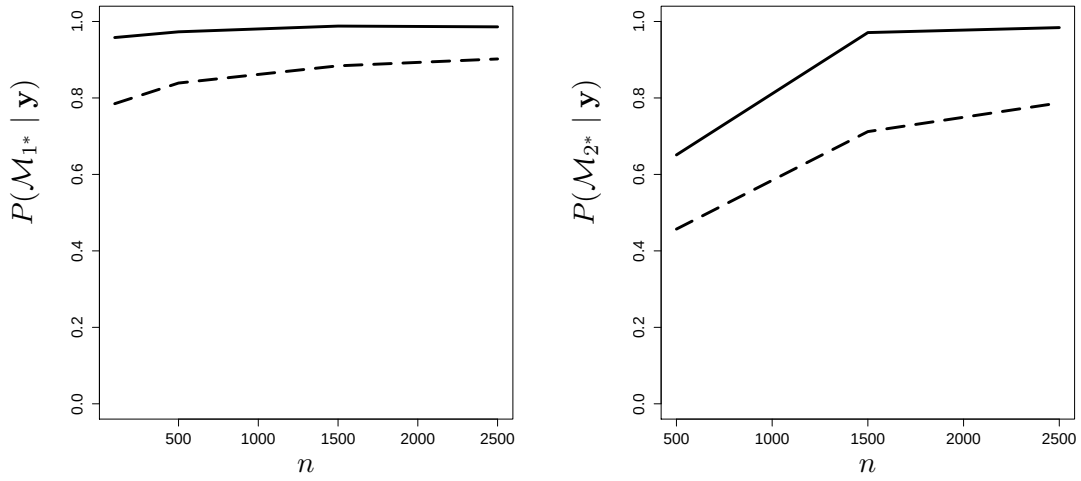
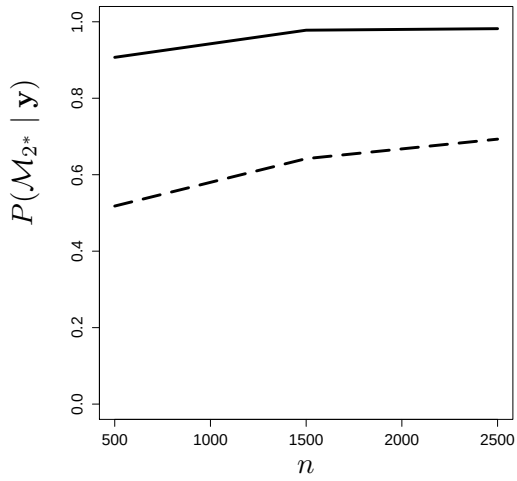


Figure 5.1: Simulation study data-generating truth.

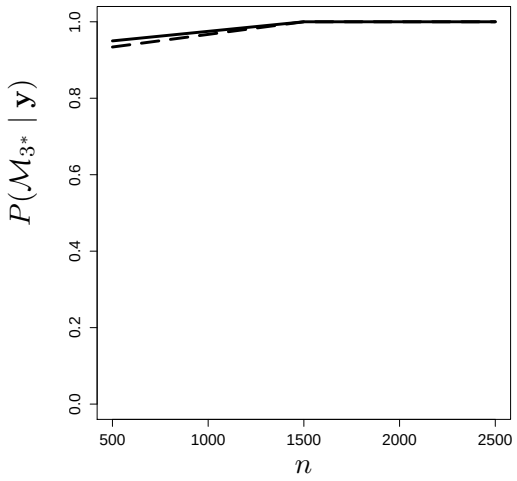


(a) Case 1

(b) Case 2

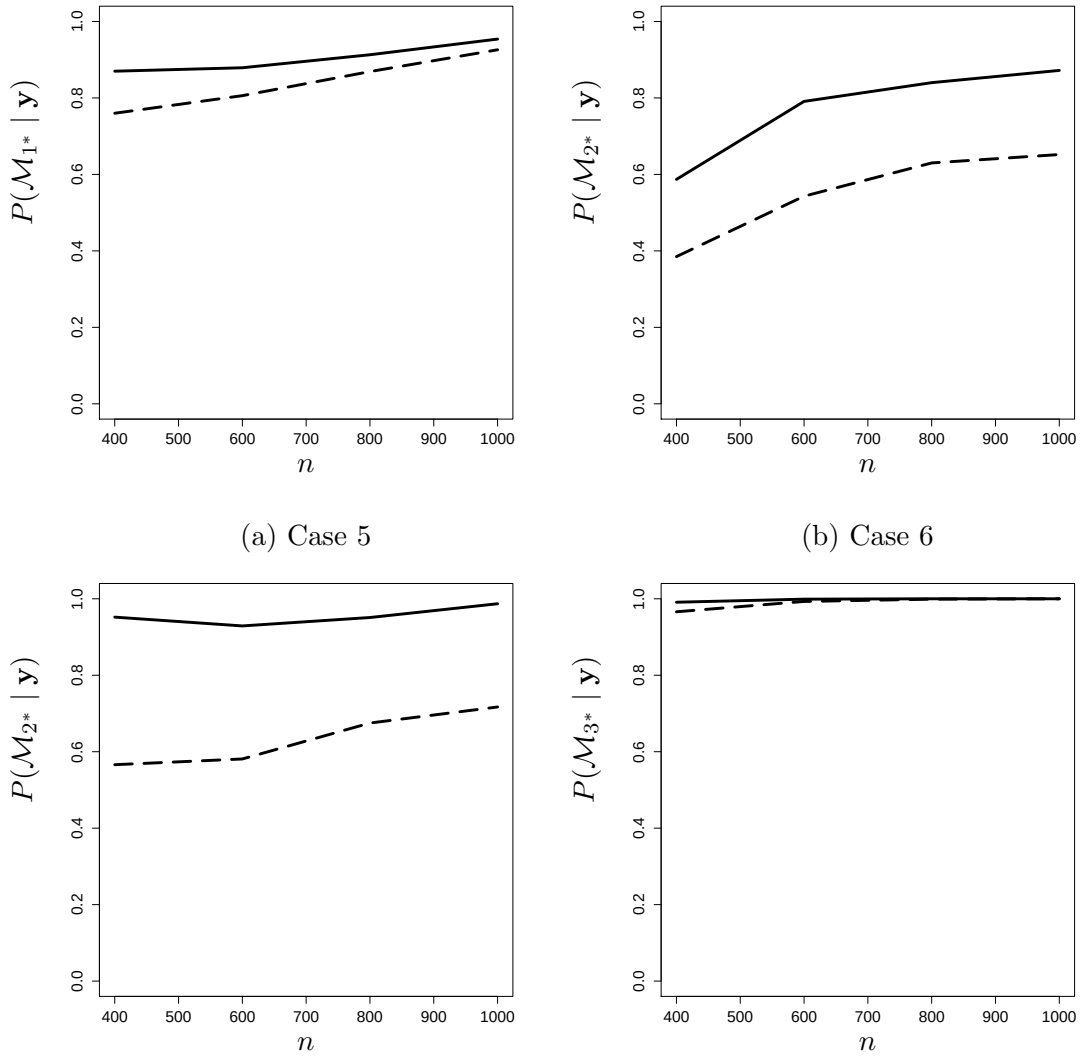


(c) Case 3



(d) Case 4

Figure 5.2: Simulation study. Univariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n for the MOM-IW (solid line) and Normal-IW (dashed line).



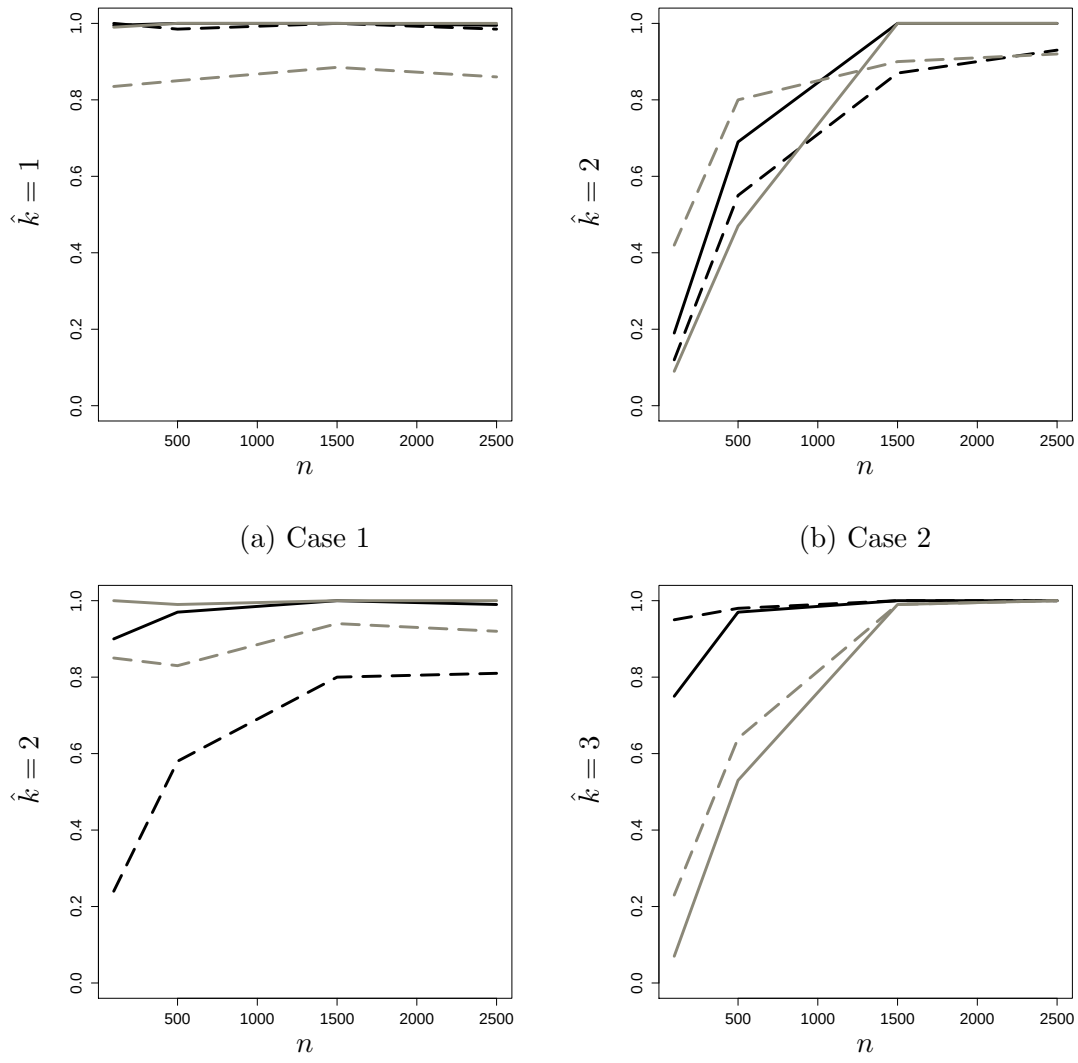
(a) Case 5

(b) Case 6

(c) Case 7

(d) Case 8

Figure 5.3: Simulation study. Bivariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n for the MOM-IW (solid line) and Normal-IW (dashed line).



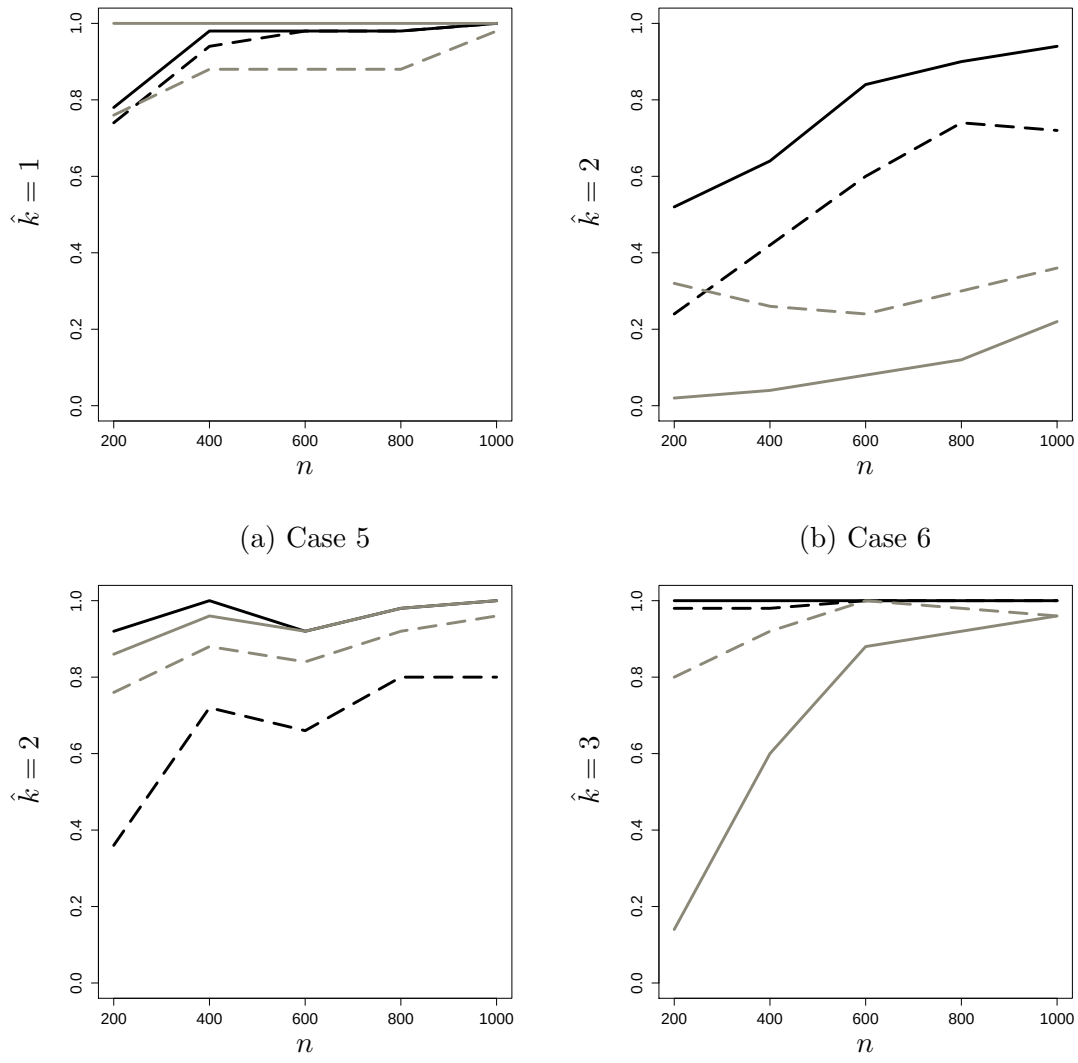
(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 5.4: Simulation study. Univariate mixtures. Proportion of correct $\hat{k} = k$ vs. n for MOM-IW (solid black), Normal-IW (dashed black), AIC (dashed gray) and BIC (solid gray).



(a) Case 5

(b) Case 6

(c) Case 7

(d) Case 8

Figure 5.5: Simulation study. Bivariate mixtures. Proportion of correct $\hat{k} = k$ vs. n for MOM-IW (solid black), Normal-IW (dashed black), AIC (dashed gray) and BIC (solid gray).

5.2 Misspecified mixtures

In practice, the data-generating density may present non-negligible departures from the assumed class. An important case we investigate here is the presence of heavy tails and asymmetries for three misspecified mixtures, which under an assumed Normal mixture likelihood may affect both the chosen k and the parameter estimates. The continuous flexible mixtures proposed by Rossell and Steel (2017) based in multivariate two piece iskew-t or dskew-t distributions capture asymmetry and heavy tails.

The iskew-t distribution considers independent observations from univariate two-piece skew-Normal distributions. The misspecified mixtures here have bivariate Student-t or iskew-t components (see Appendix C). For the three considered mixtures we generated $n = 600$ observations from $k^* = 3$, means $\boldsymbol{\mu}_1 = (-1, 1)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (6, 6)'$, a common scale matrix with elements $\sigma_{11} = \sigma_{22} = 2$ and $\sigma_{12} = \sigma_{21} = -1$ and $\eta_1 = \eta_2 = \eta_3 = 1/3$. For the bivariate Student-t components the degrees of freedom are $\upsilon_j = 4$ and for the bivariate iskew-t components are $\upsilon_j = 4$ or $\upsilon_j = 100$ (in each mixture). For the bivariate iskew-t component the j skewness parameter for variable $\mathbf{y}$ is equal to $\alpha^s_{jf} = -0.5$ which corresponds to a positive skewness. We considered up to $K = 6$ components for each of the three misspecified mixtures with either homogeneous $\Sigma_1 = \ldots = \Sigma_k$ or heterogeneous covariances, giving a total of 11 models.

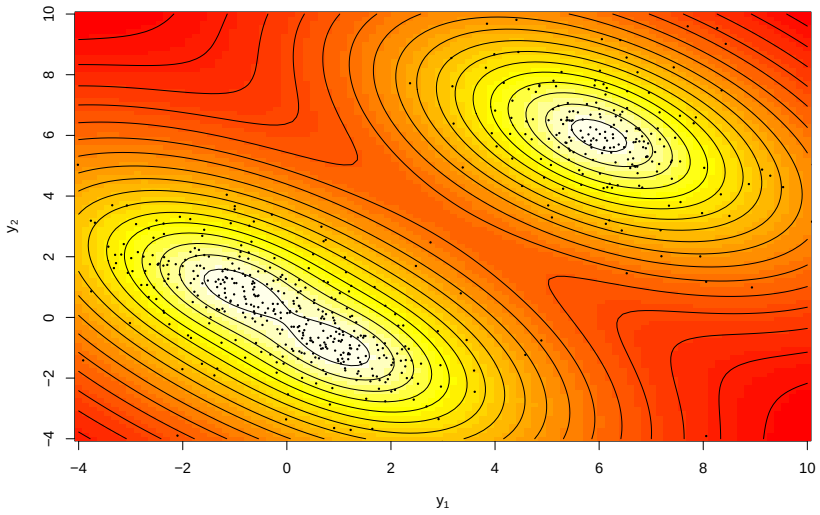


Figure 5.6: Simulated data and contour in logarithm scale for the data-generating student-T mixture with $\upsilon_j = 4$.

Table 5.2 summarizes the results. For the student-T mixture illustrated in Figure 5.6, BIC and sBIC strongly favored $\hat{k} = 4$ components with unequal covariances, AIC chose $\hat{k} = 6$ components with unequal covariances, and the Normal-IW prior placed most posterior probability on $k \in \{5, 6\}$ with common covariances.

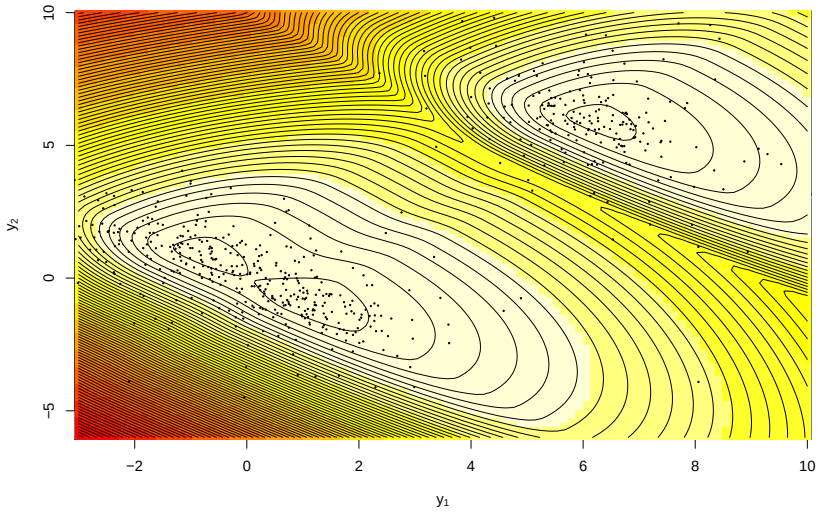


Figure 5.7: Simulated data and contour in logarithm scale for the data-generating iskew-T mixture with $\upsilon_j = 100$.

For the iskew-T mixture with $\upsilon_j = 100$ illustrated in Figure 5.7, AIC and sBIC supported $\hat{k} = 6$ components with unequal covariances, the Normal-IW prior and BIC chose $\hat{k} = 4$ and $\hat{k} = 5$ components with a common covariance, respectively. For the iskew-T mixture with $\upsilon_j = 4$ displayed in Figure 5.8, BIC and sBIC indicate $\hat{k} = 5$ components with unequal covariances. The Normal-IW prior and AIC chose $\hat{k} = 6$ components with common and unequal covariances, respectively. In contrast, our MOM-IW assigned posterior probability 1 (up to rounding) to $k = 3$ with equal covariances for the student-T mixture.

For the iskew-T mixtures with $\upsilon_j = 100$ and $\upsilon_j = 4$, the proposed MOM-IW chose $k = 3$ and $k = 5$ components with equal covariances and induced posterior probabilities of 0.863 and 0.651, respectively. To provide further insight Figures 5.9-5.11 show the component contours for $\hat{k}$ under each method, estimating $\hat{\boldsymbol{\vartheta}}_{\hat{k}}$ via maximum likelihood (AIC, BIC, sBIC) or posterior modes (Normal-IW, MOM-IW).

In Figure 5.9 the means of the three MOM-IW components matched those of the true Student-T components. The BIC and sBIC approximated the two mildly-separated components with two normals centered roughly at (0,0), whereas the AIC split the components even further. The two extra components in the Normal-IW solution essentially account for heavy tails.

In Figure 5.10 for the iskew-T mixture with $\upsilon_j = 100$ the means of the three MOM-IW components are closer to those of the true components. Additional components are suggested by AIC, BIC, sBIC and Normal-IW due to the presence of skewness in the mixture. Figure 5.11 displays how AIC, BIC, sBIC, Normal-IW and MOM-IW suggest more components than $k^* = 3$ due to the presence of both asymmetry and thick tails in the mixture and how MOM-IW induces some parsimony regardless. This example illustrates how by penalizing poorly-separated or low-weight components NLPs may induce a form of robustness to model misspecification for mixtures with heavy tails or asymmetries, although we remark that this is a finite-sample effect and would eventually vanish as $n \to \infty$. Even though this is computationally convenient we remark that in general the criterion to discard components is case-dependent and, unless carefully calibrated, the quality of the inference may suffer.

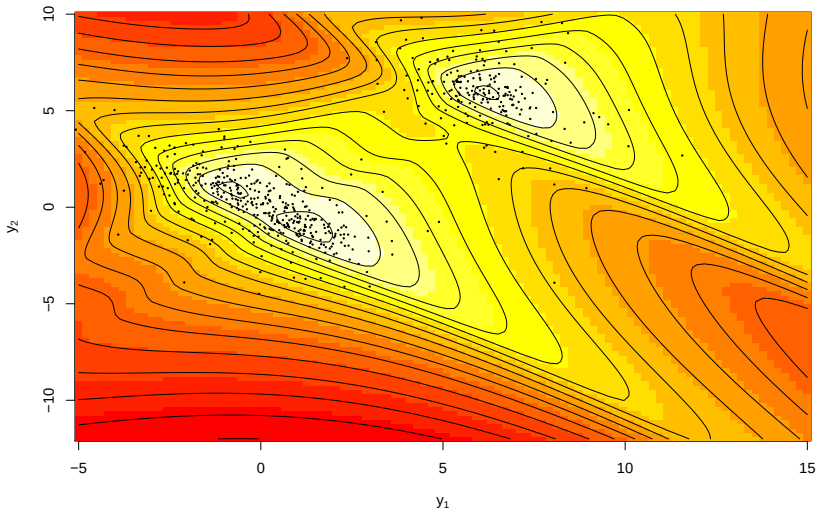
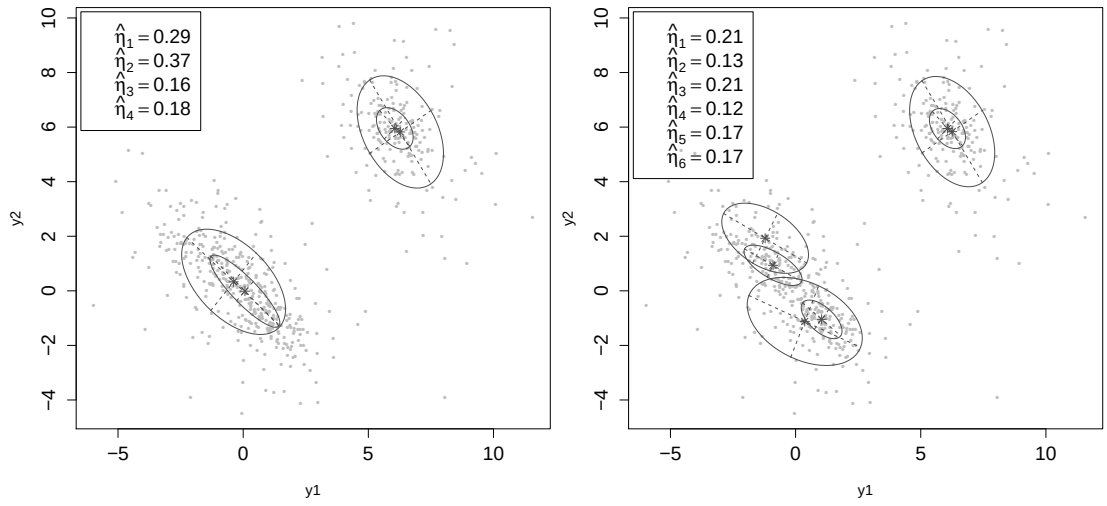


Figure 5.8: Simulated data and contour in logarithm scale for the data-generating iskew-T mixture with $\upsilon_j = 4$.

For instance, Figure 5.9 illustrates how in the student-T mixture for $K = 6$ the BIC, sBIC, AIC and the Normal-IW support 4 components by setting an estimated weight $\hat{\eta}_j$ >0.15 that would not be discarded in practice.

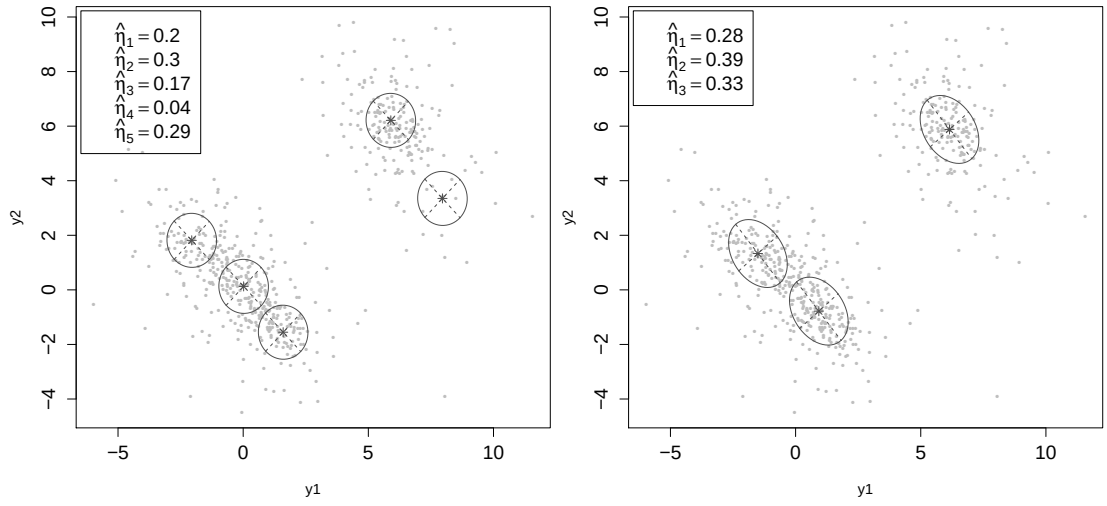
Table 5.2: Misspecified student-T mixture, iskew-T mixture with $\upsilon_j = 100$ and iskew-T mixture with $\upsilon_j = 4$. $P(\mathcal{M}_k \mid \mathbf{y})$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under Normal-IW-Dir, MOM-IW-Dir, BIC and sBIC under $\Sigma_i \neq \Sigma_j$.

			Student-T with $\upsilon_j = 4$			
		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
	1	0.000	0.000	-2992.820	-2981.828	
$\Sigma_j = \Sigma$	2	0.000	0.000	-2549.767	-2532.179	
	3	0.003	**1.000**	-2548.774	-2524.591	
	4	0.062	0.000	-2556.581	-2525.803	
	5	**0.469**	0.000	-2566.122	-2528.748	
	6	0.465	0.000	-2574.371	-2530.402	
$\Sigma_i \neq \Sigma_j$	2	0.000	0.000	-2545.129	-2520.946	-2548.942
	3	0.000	0.000	-2529.037	-2491.663	-2534.729
	4	0.000	0.000	**-2522.954**	-2472.389	**-2527.448**
	5	0.000	0.000	-2535.703	-2471.948	-2528.207
	6	0.000	0.000	-2546.878	**-2469.931**	-2529.068
			iskew-T with $\upsilon_j = 100$			
		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
	1	0.000	0.000	-2887.208	-2876.216	
$\Sigma_j = \Sigma$	2	0.000	0.000	-2269.108	-2251.520	
	3	0.148	**0.863**	-2248.125	-2223.942	
	4	**0.852**	0.137	-2250.330	-2219.551	
	5	0.000	0.000	**-2244.224**	-2206.850	
	6	0.000	0.000	-2248.776	-2204.807	
$\Sigma_i \neq \Sigma_j$	2	0.000	0.000	-2256.182	-2231.999	-2259.994
	3	0.000	0.000	-2250.169	-2212.795	-2252.693
	4	0.000	0.000	-2253.893	-2203.329	-2250.255
	5	0.000	0.000	-2251.225	-2191.983	-2244.667
	6	0.000	0.000	-2268.929	**-2187.470**	**-2244.587**
			iskew-T with $\upsilon_j = 4$			
		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
	1	0.000	0.000	-3035.295	-3024.302	
$\Sigma_j = \Sigma$	2	0.000	0.000	-2574.563	-2556.975	
	3	0.000	0.000	-2559.473	-2535.290	
	4	0.162	0.308	-2568.440	-2537.662	
	5	0.263	**0.651**	-2553.298	-2515.924	
	6	**0.576**	0.041	-2533.806	-2489.836	
$\Sigma_i \neq \Sigma_j$	2	0.000	0.000	-2572.904	-2548.721	-2576.716
	3	0.000	0.000	-2539.473	-2502.099	-2534.317
	4	0.000	0.000	-2531.996	-2481.431	-2526.147
	5	0.000	0.000	**-2531.954**	-2468.199	**-2520.697**
	6	0.000	0.000	-2543.217	**-2466.271**	-2523.186



(a) BIC/sBIC

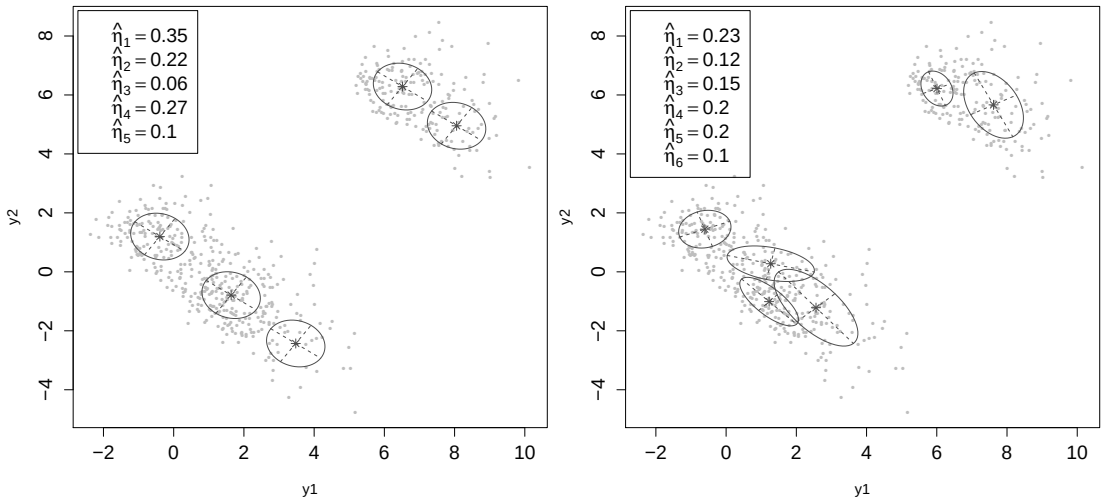
(b) AIC



(c) Normal-IW

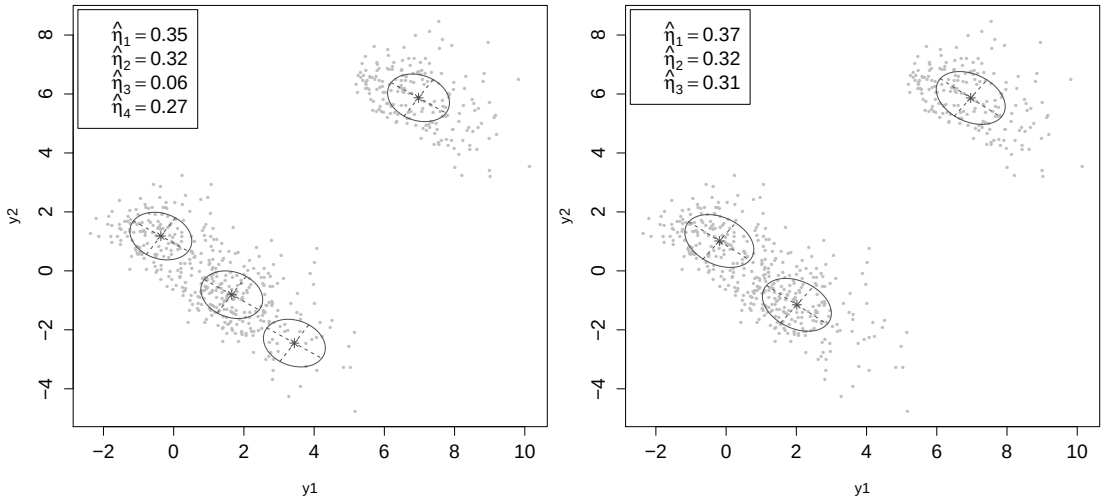
(d) MOM-IW

Figure 5.9: Misspecified Student-T mixture. Estimated contours for (a) BIC/sBIC (top left), (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data.



(a) BIC

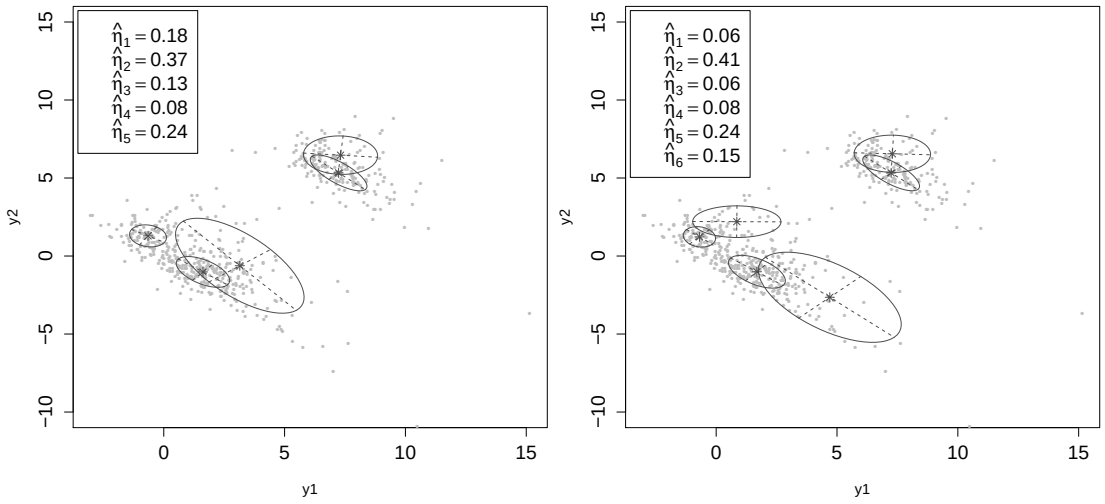
(b) AIC/sBIC



(c) Normal-IW

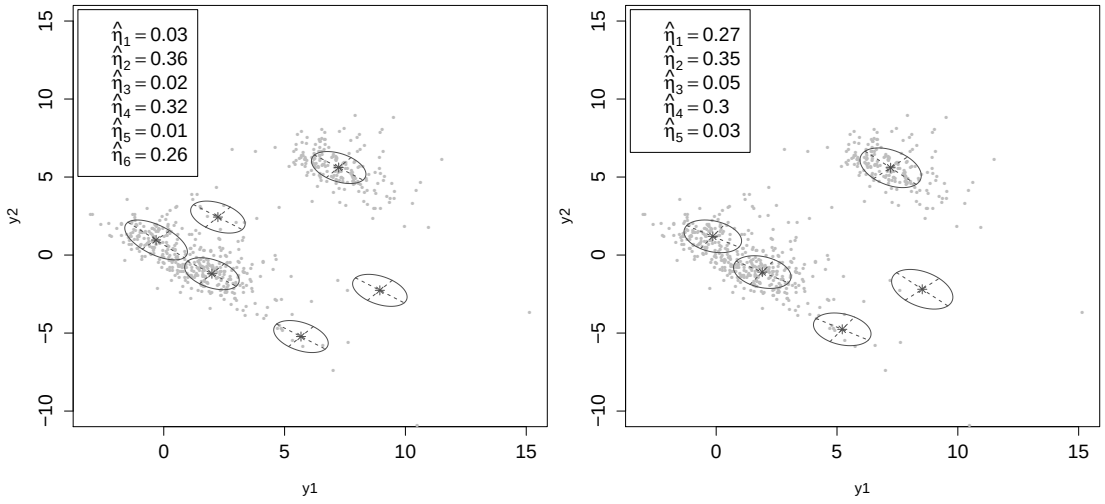
(d) MOM-IW

Figure 5.10: Misspecified iskew-T mixture with $\upsilon_j = 100$. Estimated contours for (a) BIC (top left), (b) AIC/sBIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data.



(a) BIC/sBIC

(b) AIC



(c) Normal-IW

(d) MOM-IW

Figure 5.11: Misspecified iskew-T mixture with $\upsilon_j = 4$. Estimated contours for (a) BIC/sBIC (top left), (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the simulated data.

5.3 Binomial mixture

To assess the performance of MOM-Beta (default $a = 1/2$, $g = 6.93$) and Beta(1,1) priors, the BIC and sBIC we reproduced the Binomial mixture example used by Drton and Plummer (2017) to illustrate the sBIC. We generated 200 data sets of sample sizes $n = 50$, 200 and 500 from a $k^* = 4$ component Binomial mixture with $L_{if} = 30$ trials for all $i = 1, \ldots, n$, equal component weights $\eta_j = 1/4$ and component-specific success probabilities $\theta_j = j/5$ for $j = 1, \ldots, 4$. To compute the sBIC we considered the two different bounds for the real canonical threshold namely $\lambda \leq \frac{1}{2}(k + j - 1)$ and $\lambda \leq \frac{1}{4}(j + 3k) - \frac{1}{2}$ proposed by Drton and Plummer (2017). These two sBIC versions are denoted as $\overline{\text{sBIC}}$ and $\overline{\text{sBIC}}_{05}$, respectively.

Figure 5.12 shows the results. The two sBIC versions ameliorated the BIC's overpenalization as reported in Drton and Plummer (2017), whereas the Beta prior often returned too many components.

The proportion of correct model selections was generally highest for the MOM-Beta, particularly for smaller n (roughly 50% of the simulations when $n = 50$, relative to the 25% for $\overline{\text{sBIC}}_{05}$).

5.4 Sensitivity to the prior

We perform a sensitivity analysis for the choice of q of the Dir($\boldsymbol{\eta}$;q) prior in Normal mixtures. Using the posterior expected number of components given by $E(k \mid \mathbf{y}) = P(\mathcal{M}_1 \mid \mathbf{y}) + 2P(\mathcal{M}_2 \mid \mathbf{y}) + 3P(\mathcal{M}_3 \mid \mathbf{y})$ we set $q = p + 1$ as the default prior specification of q (Section 3.2) and compare with the specification of $q = 4$ and $q = 16.5$ for univariate and bivariate Normal mixtures as suggested in Frühwirth-Schnatter (2006). We also assess the sensitivity of choosing g for some alternative prior parameter settings for MOM-IW and MOM-Beta priors.

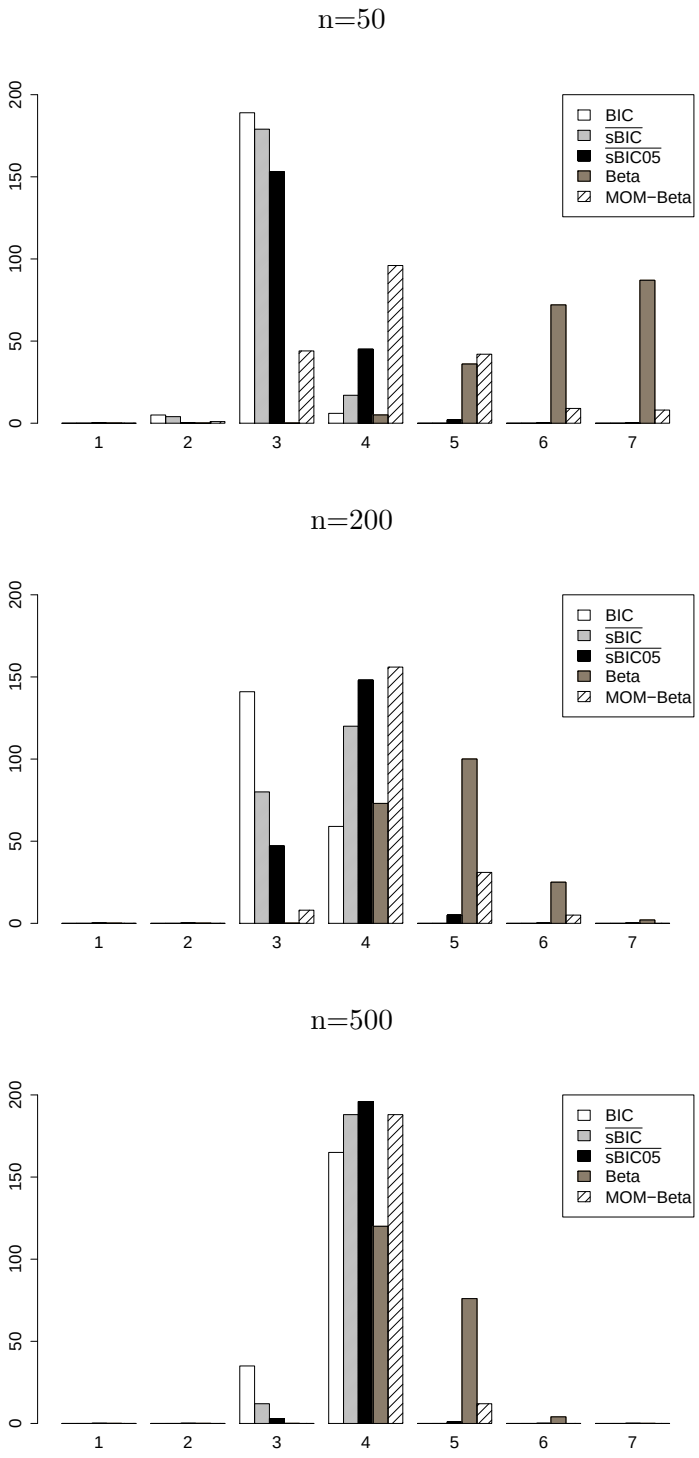


Figure 5.12: Binomial mixture. Frequencies of $\hat{k}$ for BIC, $\overline{\text{sBIC}}$, $\overline{\text{sBIC}}_{05}$, Beta and MOM-Beta. Results from 200 data sets with $n = 50$, 200 and 500, $L_{if} = 30$ and $k^* = 4$.

5.4.1 Sensitivity to choosing q

Regarding the univariate Normal mixtures in Cases 1-4, the four panels in Figure 5.13 show $E(k \mid \mathbf{y})$ for the alternative prior specification $q = 2$ and $P(\kappa < 4) = 0.05$. The four panels in Figure 5.14 show analogous results for $q = 4$ and $P(\kappa < 4) = 0.05$, showing that the findings are fairly robust to mild deviations from our default value of q.

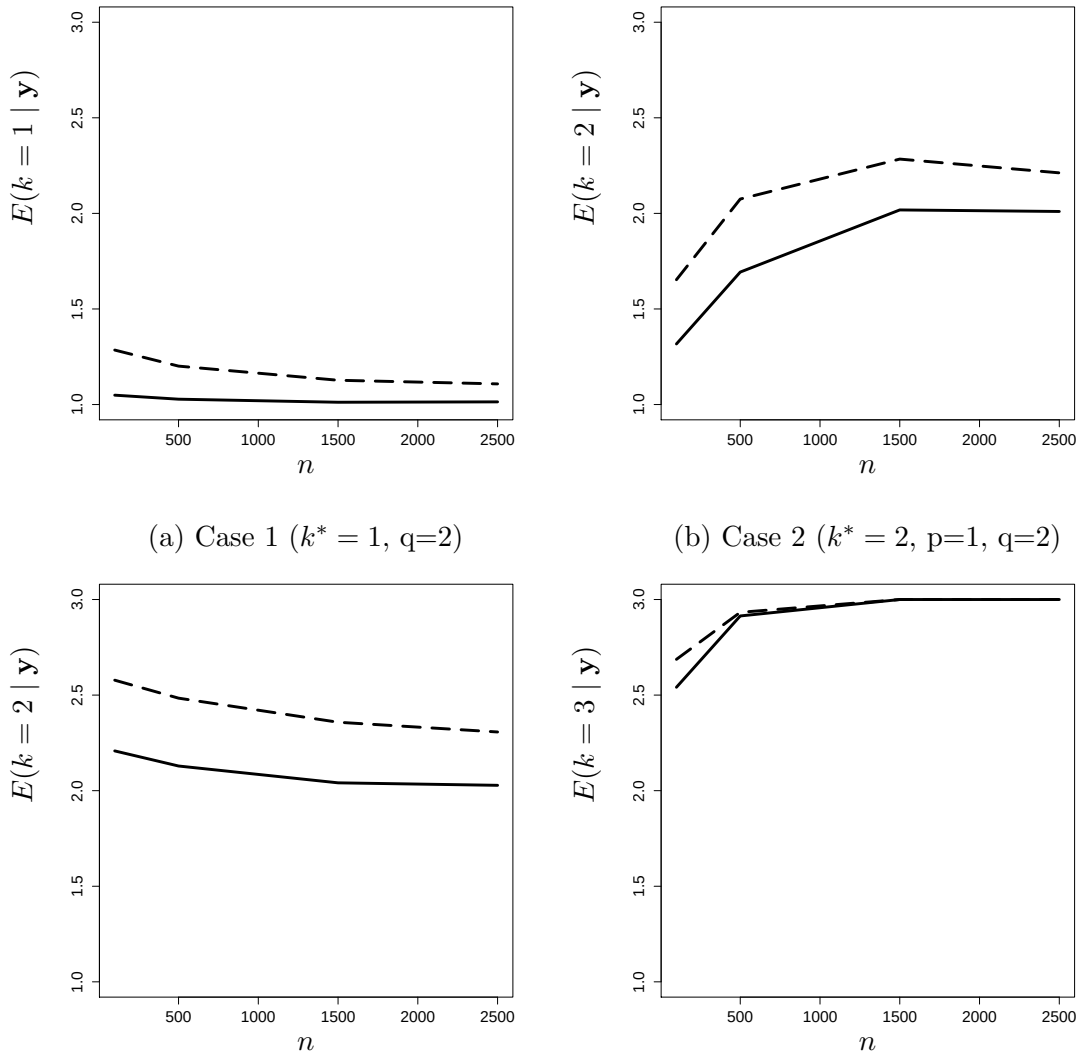
Regarding the bivariate Normal mixtures in Cases 5-8, the four panels in Figure 5.15 show $E(k \mid \mathbf{y})$ for $q = 3$ and $P(\kappa < 4) = 0.05$. The four panels in Figure 5.16 show the same results for $q = 16.5$ (a value recommended in Frühwirth-Schnatter (2006) and Mengersen et al. (2011), Chapter 10) and $P(\kappa < 4) = 0.05$, showing again that the findings are fairly robust to mild deviations from our recommended prior setting.

5.4.2 Sensitivity to choosing g for MOM-IW priors

To study the sensitivity to the prior elicitation of g, Figures 5.17-5.18 show the average posterior probability $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ for Cases 1-8 with $P(\kappa < 4) = 0.1$ and q set as in Figures 5.2-5.3. Although the results are largely similar to those in Figures 5.2-5.3, the benefits in parsimony enforcement are somewhat reduced in some situations (*e.g.* Case 5), indicating that $P(\kappa < 4 \mid g, \mathcal{M}_K) = 0.05$ may be slightly preferable to 0.1 to achieve a better balance between parsimony and detection power.

5.4.3 Sensitivity to choosing g for MOM-Beta priors

To assess sensitivity Figure 5.19 shows the results for alternative prior parameter settings $g = 7.05$, $g = 16.09$ and $g = 29.99$ discussed in Section 3.2. While the performance remains competitive, these larger g result in more informative priors that adversely affect inference, reinforcing our default recommendation $g = 7.05$.



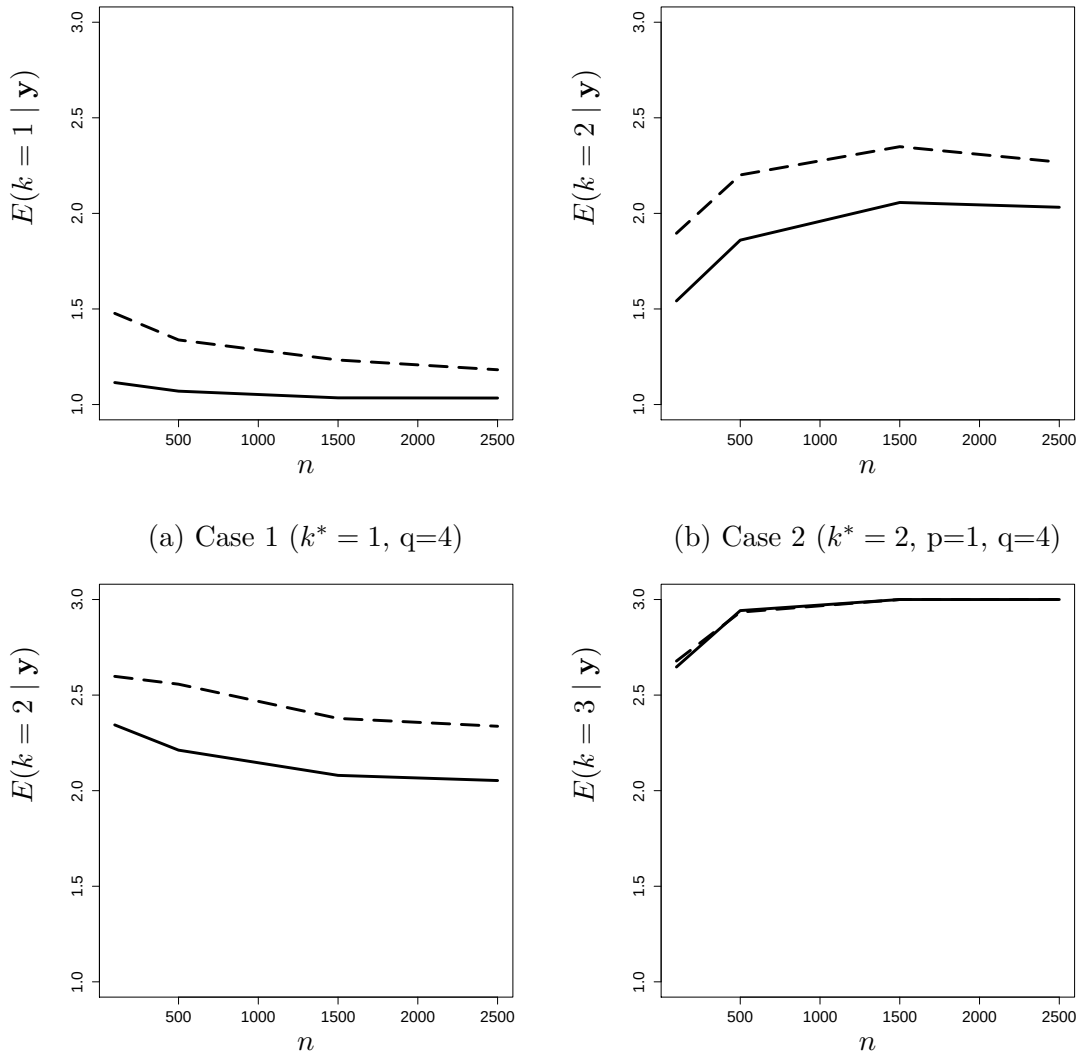
(a) Case 1 ($k^* = 1$, q=2)

(b) Case 2 ($k^* = 2$, p=1, q=2)

(c) Case 3 ($k^* = 2$, p=1, q=2)

(d) Case 4 ($k^* = 3$, p=1, q=2)

Figure 5.13: Simulation study. Univariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = p + 1$ for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).



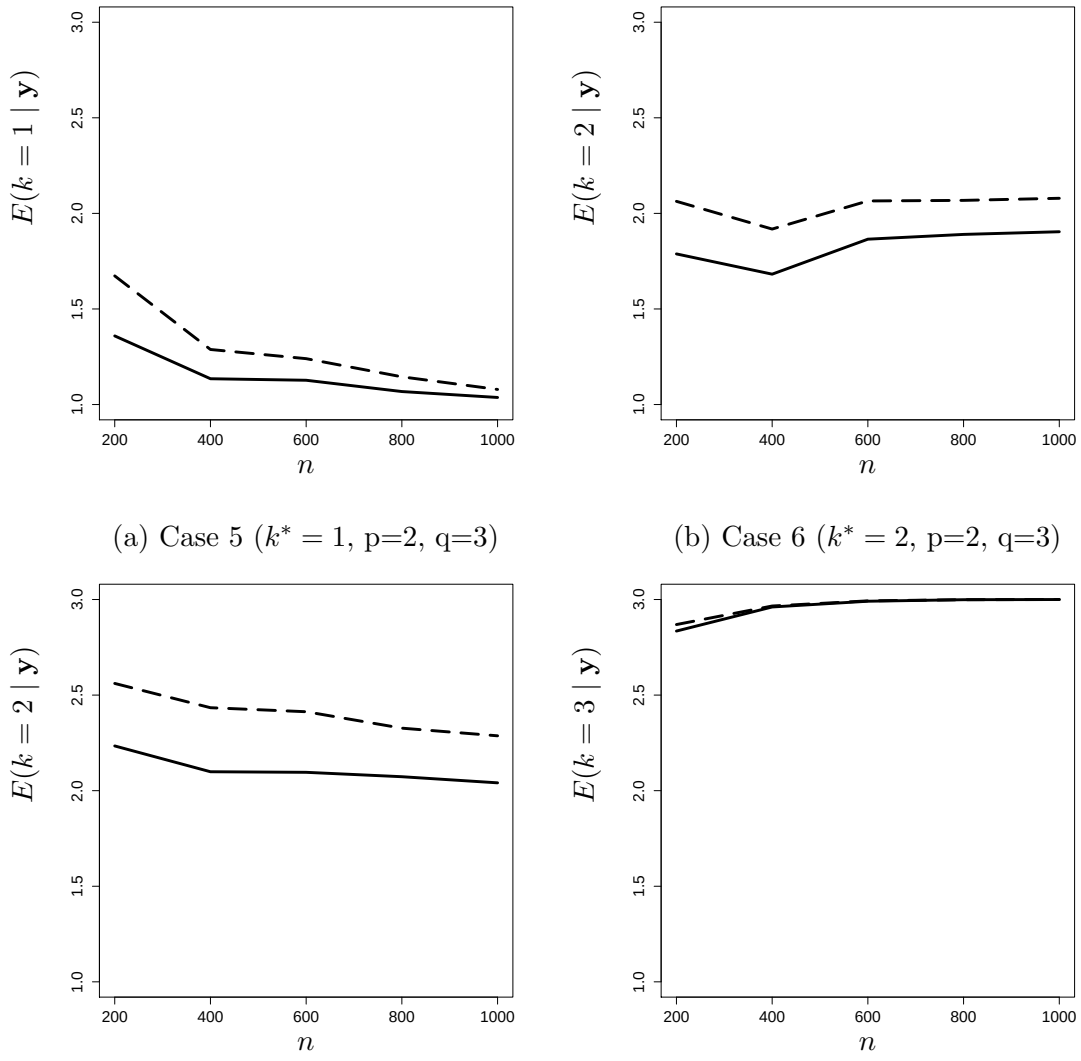
(a) Case 1 ($k^* = 1$, q=4)

(b) Case 2 ($k^* = 2$, p=1, q=4)

(c) Case 3 ($k^* = 2$, p=1, q=4)

(d) Case 4 ($k^* = 3$, p=1, q=4)

Figure 5.14: Simulation study. Univariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = 4$ as recommended by Frühwirth-Schnatter (2006) for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).



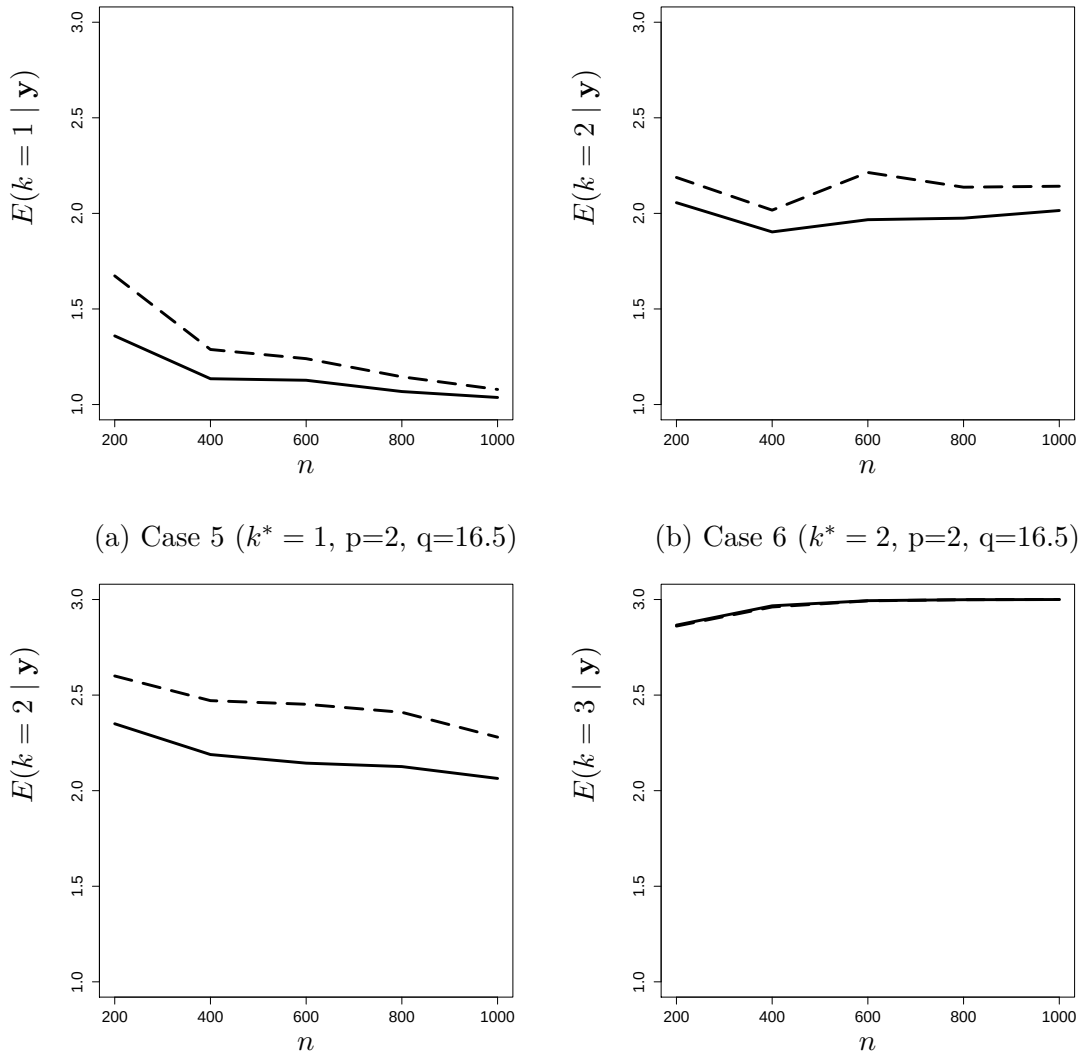
(a) Case 5 ($k^* = 1$, p=2, q=3)

(b) Case 6 ($k^* = 2$, p=2, q=3)

(c) Case 7 ($k^* = 2$, p=2, q=3)

(d) Case 8 ($k^* = 3$, p=2, q=3)

Figure 5.15: Simulation study. Bivariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = p + 1$ for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).



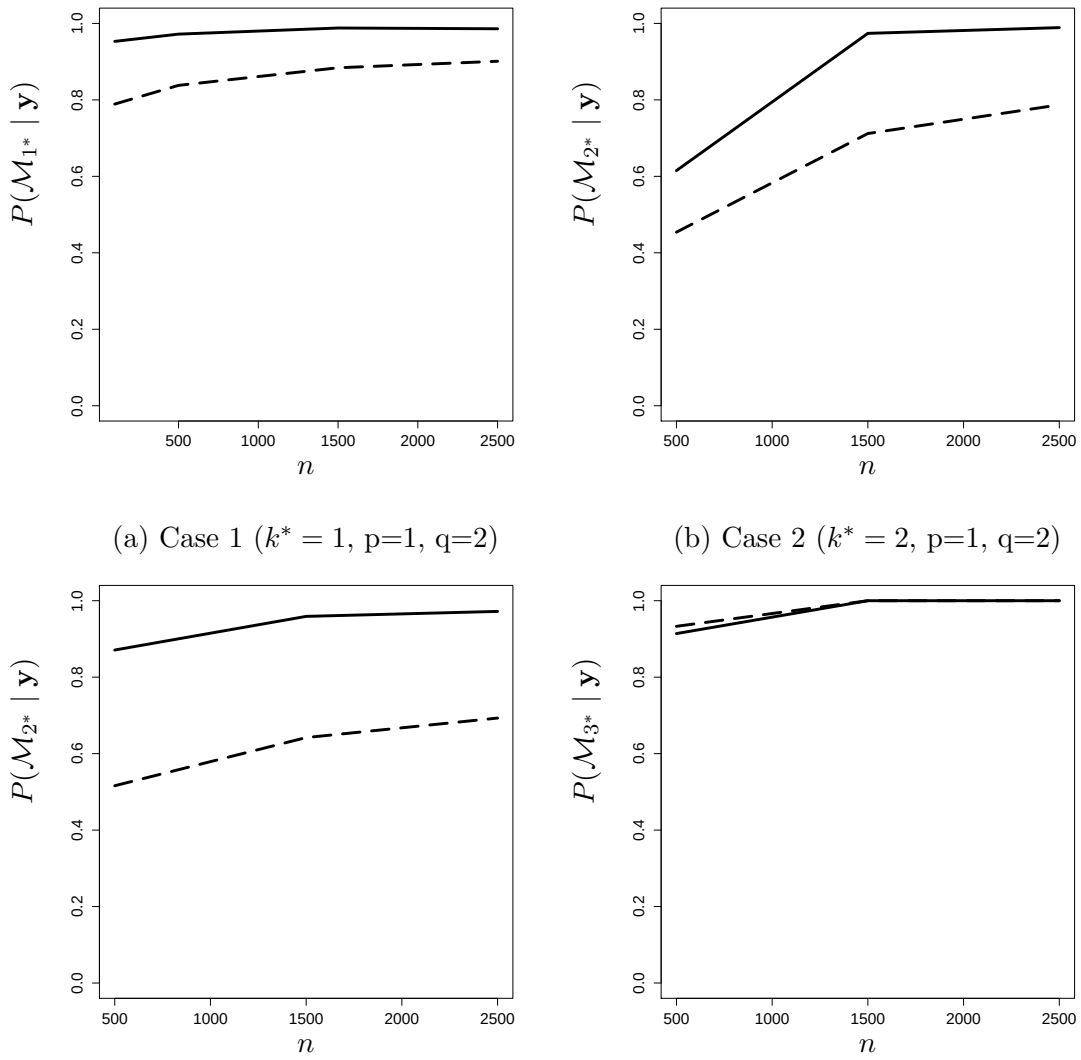
(a) Case 5 ($k^* = 1$, p=2, q=16.5)

(b) Case 6 ($k^* = 2$, p=2, q=16.5)

(c) Case 7 ($k^* = 2$, p=2, q=16.5)

(d) Case 8 ($k^* = 3$, p=2, q=16.5)

Figure 5.16: Simulation study. Bivariate mixtures. Posterior expected model size $E(k \mid \mathbf{y})$ versus n with $q = 16.5$ as recommended by Frühwirth-Schnatter (2006) for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).



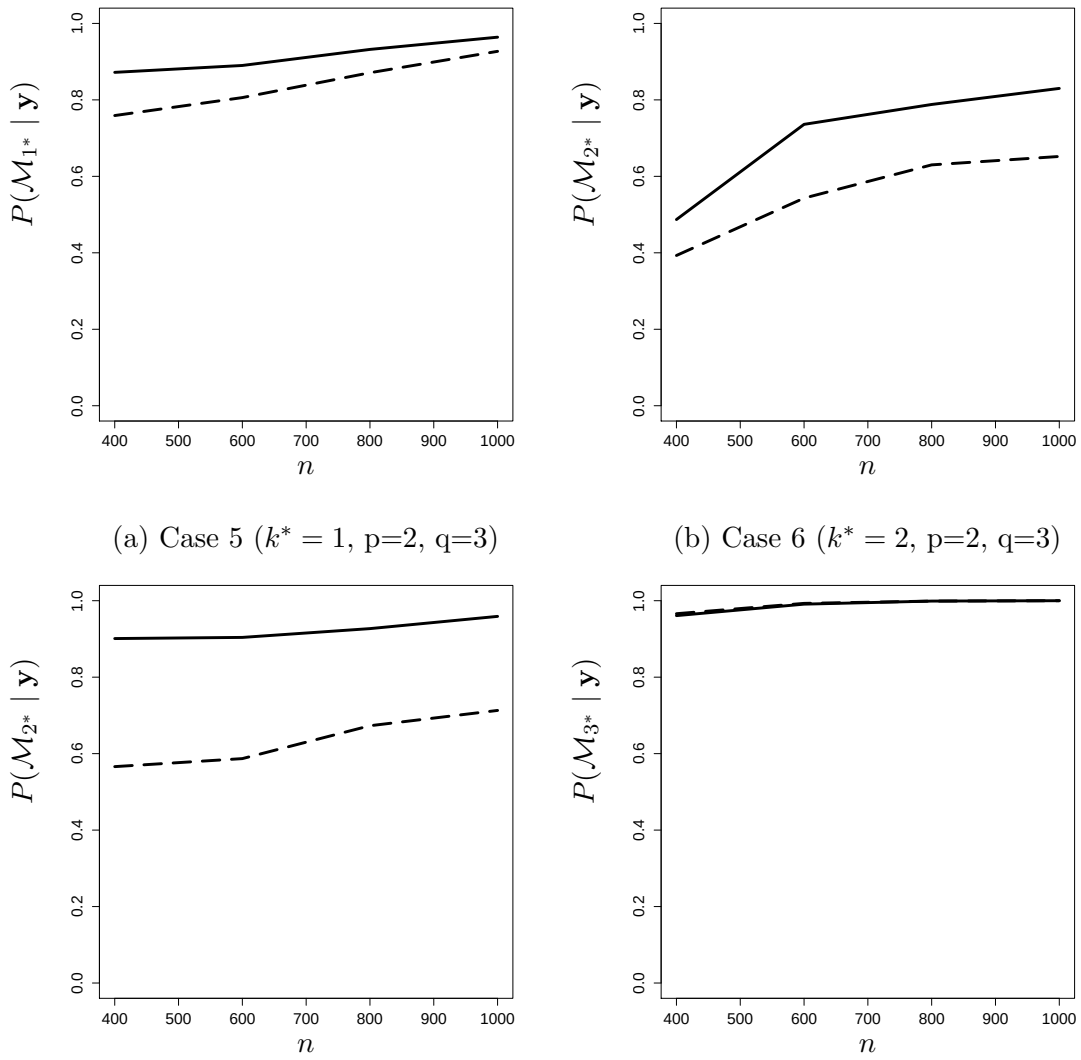
(a) Case 1 ($k^* = 1$, p=1, q=2)

(b) Case 2 ($k^* = 2$, p=1, q=2)

(c) Case 3 ($k^* = 2$, p=1, q=2)

(d) Case 4 ($k^* = 3$, p=1, q=2)

Figure 5.17: Simulation study. Univariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n under $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$ for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).



(a) Case 5 ($k^* = 1$, p=2, q=3)

(b) Case 6 ($k^* = 2$, p=2, q=3)

(c) Case 7 ($k^* = 2$, p=2, q=3)

(d) Case 8 ($k^* = 3$, p=2, q=3)

Figure 5.18: Simulation study. Bivariate mixtures. $P(\mathcal{M}_{k^*} \mid \mathbf{y})$ versus n under $P(\kappa < 4 \mid \mathcal{M}_k) = 0.1$ for the MOM-IW-Dir (solid line) and Normal-IW-Dir (dashed line).

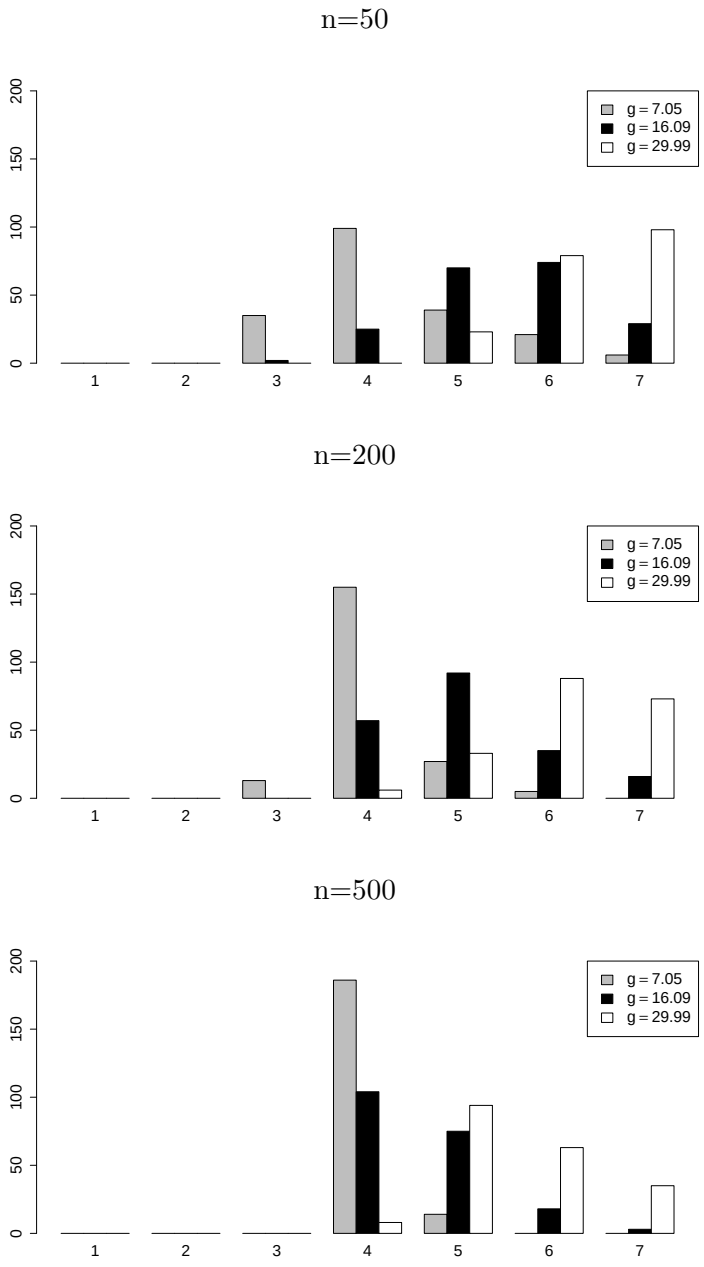


Figure 5.19: Binomial mixture. Frequencies of $\hat{k}$ for MOM-Beta for $g = 7.05$, $g = 16.09$ and $g = 29.99$ with $q = 2$. Results from 200 data sets with $n = 50$, $n = 200$ and $n = 500$, $L_{if} = 30$ and $k^* = 4$.

5.5 An illustration of computations under product of Binomial mixtures

We illustrate some computational issues and diagnostics related to posterior multimodality, the EM and MCMC algorithms in product Binomial mixtures. We considered a simulation with $k^* = 4$ components, $n = 500$, $p = 8$ variables and equal component weights $\eta_1^* = \eta_2^* = \eta_3^* = \eta_4^* = 1/4$. Each component had two large success probabilities $\theta_{jf}^* = 0.32$ whereas the remaining probabilities were small (0.04 and 0.08), specifically

$$\eta = (0.25, 0.25, 0.25, 0.25); \qquad \boldsymbol{\theta} = \begin{pmatrix} 0.32 & 0.04 & 0.04 & 0.04 \\ 0.32 & 0.08 & 0.08 & 0.08 \\ 0.04 & 0.32 & 0.04 & 0.04 \\ 0.08 & 0.32 & 0.08 & 0.08 \\ 0.04 & 0.04 & 0.32 & 0.04 \\ 0.08 & 0.08 & 0.32 & 0.08 \\ 0.04 & 0.04 & 0.04 & 0.32 \\ 0.08 & 0.08 & 0.08 & 0.32 \end{pmatrix}. \tag{5.5.1}$$

The default MOM-Beta prior parameters are $g = 2.6$ and $q = 2$ (Section 3.2). Although our EM algorithm is guaranteed to increase the log-posterior at each iteration, in practice there are potential issues with slow convergence or reaching local maxima/saddlepoints. To address this in our implementation we run the EM algorithm (Algorithm 4) from 30 different random starting values and keep the estimate achieving the highest log-posterior value. We found that for 29 of the 30 starting values the algorithm converges to a global maximum with a log-posterior value of -9906.67. The obtained estimates were fairly close to the simulation truth, specifically

$$\hat{\boldsymbol{\eta}} = (0.28, 0.26, 0.24, 0.22); \qquad \hat{\boldsymbol{\theta}} = \begin{pmatrix} 0.34 & 0.05 & 0.04 & 0.04 \\ 0.28 & 0.07 & 0.08 & 0.07 \\ 0.04 & 0.31 & 0.04 & 0.05 \\ 0.08 & 0.31 & 0.08 & 0.08 \\ 0.05 & 0.05 & 0.35 & 0.06 \\ 0.09 & 0.08 & 0.31 & 0.08 \\ 0.03 & 0.04 & 0.04 & 0.33 \\ 0.09 & 0.09 & 0.07 & 0.30 \end{pmatrix} \tag{5.5.2}$$

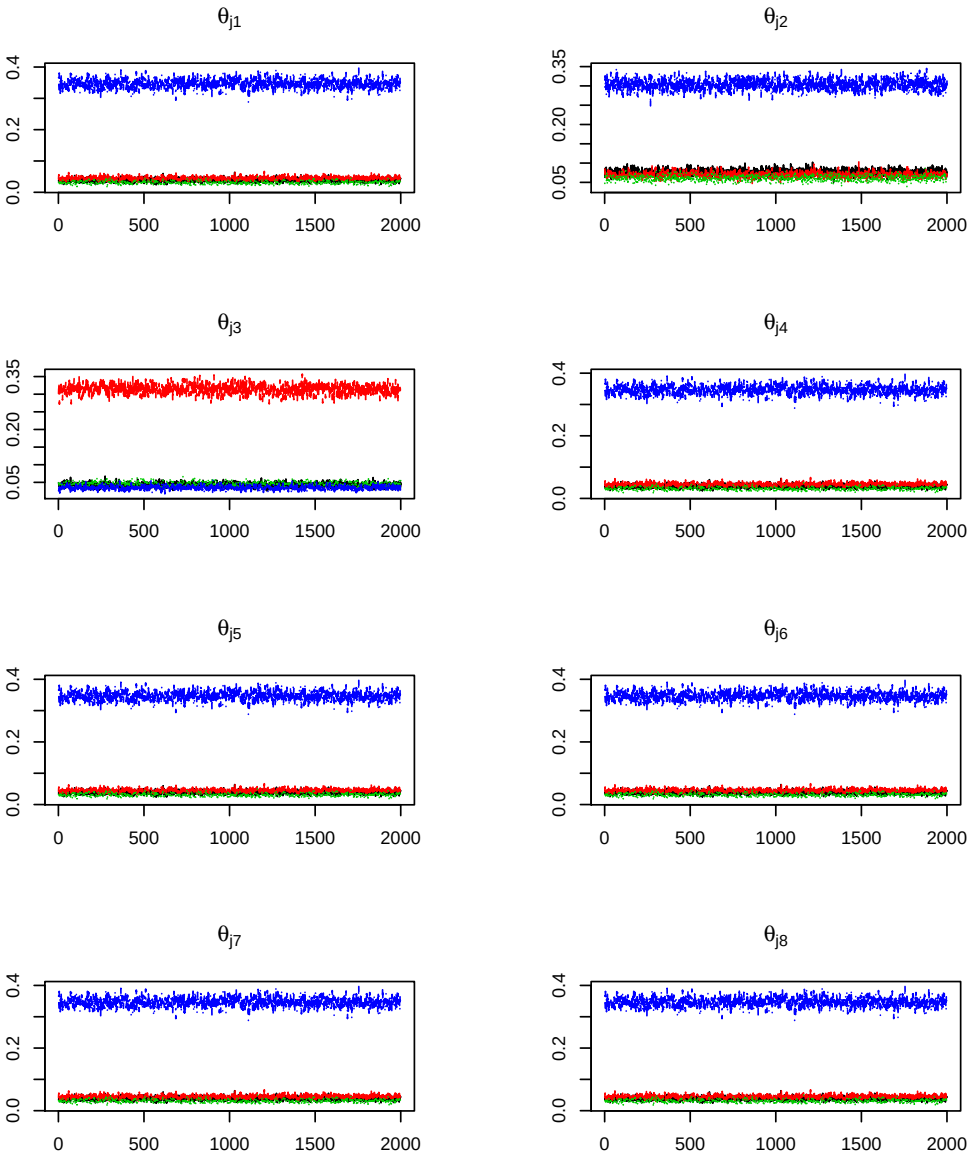


Figure 5.20: Product Binomial simulation. MCMC trace plots for θ_{jf} corresponding to components $j = 1, ..., 4$ and variables $f = 1, \ldots, 8$. The colours in the trace plots indicate the different components.

We also studied the ability of the BIC, AIC, and Beta and MOM-Beta priors to recover $k^* = 4$, finding that all except for the AIC returned the correct value (Table 5.3). Recall that the posterior probabilities require estimating the integrated likelihood, for which in turn we run an MCMC algorithm.

Because the proposed algorithm to estimate the integrate likelihood requires the MCMC to converge we assess practical MCMC convergence Figure 5.20 provides trace plots for 2,000 iterations targetting $p(\boldsymbol{\vartheta}_4 \mid \mathbf{y}, \mathcal{M}_4)$ after a burn period of 1,000. The plots do not reveal issues with the mixing.

Table 5.3: Product Binomial simulation. $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, \ldots, 6\}$ and $k^* = 4$ under Beta and MOM-Beta priors, BIC and AIC.

	Beta	MOM-Beta	BIC	AIC
k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$		
1	0.000	0.000	-22702.00	-22668.29
2	0.000	0.000	-21569.65	-21498.00
3	0.000	0.000	-20782.58	-20673.00
4	**1.000**	**1.000**	**-20051.63**	-19904.11
5	0.000	0.000	-20074.65	-19889.21
6	0.000	0.000	-20099.17	**-19875.80**

In this chapter we developed different simulation studies showing that the NLP prior has a better performance than the LP prior for continuos and categorical data. For the misspecified example, the MOM-IW prior selected the true number of components even in the presence of heavy tails or skewness thus motivating in future work the use of NLPs in flexible mixtures to select the number of components. For the product of Binomial mixtures we illustrated the maximization of parameters a posteriori and the sampling of parameters via their full conditionals. Therefore, for future extensions we could consider performing collapsed Gibbs Sampling which has been recently implemented to latent block modeling (Wyse and Friel (2012)) to improve the computational time.

Chapter 6

Computationally-fast alternatives

In this chapter we study two computationally-fast alternatives to ameliorate the computational cost of computing the integrated likelihood. In Section 6.1 we explore a computationally-fast criterion to select the number of components motivated by our MOM prior and the latent cluster indicators. According to the simulation study, the criterion performed well in univariate Normal and Binomial mixtures but showed a poor behavior for bivariate Normal mixtures. This may be because the penalty in the criterion needs to be calibrated for bivariate or higher dimensions. Therefore, further study is required before one could recommend this criterion for general use. In Section 6.2 we introduce a new computational strategy that gives a direct connection between cluster occupancies and Bayes factors with the advantage that Bayes factors allow for more general model comparisons (for instance equal vs unequal covariances in Normal mixtures). This new strategy seems to be promising and may have advantages with respect to overfitted mixtures avoiding the specification of case-specific cutoff values for selecting the number of components.

6.1 Exploration of non-local model selection criteria

The goal is to find an alternative to computing the integrated likelihood $p(\mathbf{y}|\mathcal{M}_k)$. Let $\mathbf{z}$ be the latent cluster indicators. For any given value of $\mathbf{z}$, from Bayes theorem we have

$$p^L(\mathbf{y}|\mathcal{M}_k) = \frac{p^L(\mathbf{y}, \mathbf{z}|\mathcal{M}_k)}{p^L(\mathbf{z}|\mathbf{y}, \mathcal{M}_k)} = \frac{p^L(\mathbf{y}|\mathbf{z}, \mathcal{M}_k)p^L(\mathbf{z}|\mathcal{M}_k)}{p^L(\mathbf{z}|\mathbf{y}, \mathcal{M}_k)}, \tag{6.1.1}$$

in (6.1.1) $p^L(\mathbf{y}|\mathbf{z}, \mathcal{M}_k)$ has a closed-form for many common mixtures and can be extended to mixtures of latent Gaussian distributions, specifically $p^L(\mathbf{y}|\mathbf{z}, \mathcal{M}_k) = \int p(\mathbf{y}|\boldsymbol{\zeta}_k, \mathbf{z}, \mathcal{M}_k) p^L(\boldsymbol{\zeta}_k|\mathbf{z}, \mathcal{M}_k) d\boldsymbol{\zeta}_k$, where $\boldsymbol{\zeta}_k = (\boldsymbol{\omega}_k, \boldsymbol{\upsilon}_k, \boldsymbol{\alpha}_k)$ are latent variables such that $p(\mathbf{y}|\boldsymbol{\zeta}_k, \mathbf{z}, \mathcal{M}_k)$ is a Normal mixture. To approximate $p^L(\mathbf{y}|\mathbf{z}, \mathcal{M}_k)$, we can take samples $m = 1, ..., M$ from the posterior of $\boldsymbol{\zeta}_k^{(m)}$ given the cluster indicators $\mathbf{z}$ and then average $p^L(\mathbf{y}|\boldsymbol{\zeta}_k^{(m)}, \mathbf{z}, \mathcal{M}_k)$. The prior probability of the cluster configuration $\mathbf{z}$ has closed-form under a $\boldsymbol{\eta} \sim \text{Dir}(q)$ prior

$$p^L(\mathbf{z}|\mathcal{M}_k) = \frac{\Gamma(kq) \prod_{j=1}^k \Gamma((\sum_i z_i = j) + q)}{\Gamma(q)^k \Gamma(n + kq)}, \tag{6.1.2}$$

The denominator in (6.1.1) requires evaluating a sum over the k^n elements in $\mathbb{Z}$, which is computationally prohibitive. Alternatively, noting that $p^L(\mathbf{z}|\mathbf{y}, \mathcal{M}_k)$ measures the posterior certainty on cluster configuration $\mathbf{z}$, we could set $\mathbf{z}$ to the most probable cluster and replace the denominator in (6.1.1) by another measure of posterior concentration, such as the entropy estimated

$$\text{EN}(\hat{\boldsymbol{\vartheta}}|\mathbf{y}, \mathcal{M}_k) = -\sum_{i=1}^n \sum_{j=1}^k \left(\frac{\hat{\eta}_j p(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^k \hat{\eta}_j p(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_j} \right) \log \left(\frac{\hat{\eta}_j p(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^k \hat{\eta}_j p(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_j} \right), \tag{6.1.3}$$

where $\hat{\boldsymbol{\vartheta}}$ is either the MLE or MAP. If the mixture components are well separated $\text{EN}(\hat{\boldsymbol{\vartheta}}|\mathbf{y}, \mathcal{M}_k)$ will be close to zero, otherwise $\text{EN}(\hat{\boldsymbol{\vartheta}}|\mathbf{y}, \mathcal{M}_k)$ will have a large value.

The entropy estimated is considered as a penalty term in the classification likelihood information criterion proposed in Biernacki and Govaert (1997) and revisited in Biernacki et al. (2000) to define the integrated classification likelihood criterion. We explore the following criterion to choose the number of mixture components

$$\text{D}_k = \max(\log(p(\mathbf{y}|\hat{\mathbf{z}}, \mathcal{M}_k)) + \log(p(\hat{\mathbf{z}}|\mathbf{y}, \mathcal{M}_k)) + \text{EN}(\hat{\boldsymbol{\vartheta}}|\mathbf{y}) + \log(\prod_{1 \leq i < j \leq k} d(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j)), \tag{6.1.4}$$

where $\hat{\mathbf{z}}$ and $\hat{\theta}_j$ are estimated using either MCMC samples or EM algorithms. Although D_k in general does not correspond asymptotically to $\log(p(\mathbf{y}|\mathbf{z}, \mathcal{M}_k))$, we conducted an exploratory analysis to assess whether it may provide a practical and scalable strategy. We explored the performance of (6.1.4) using the univariate mixtures in Cases 1 to 4 of Section 5.1. We simulated 100 data sets with samples sizes of $n = 250$, $n = 500$ and $n = 1000$ and computed the relative frequency for the selected model using (6.1.4) and MOM-IW priors.

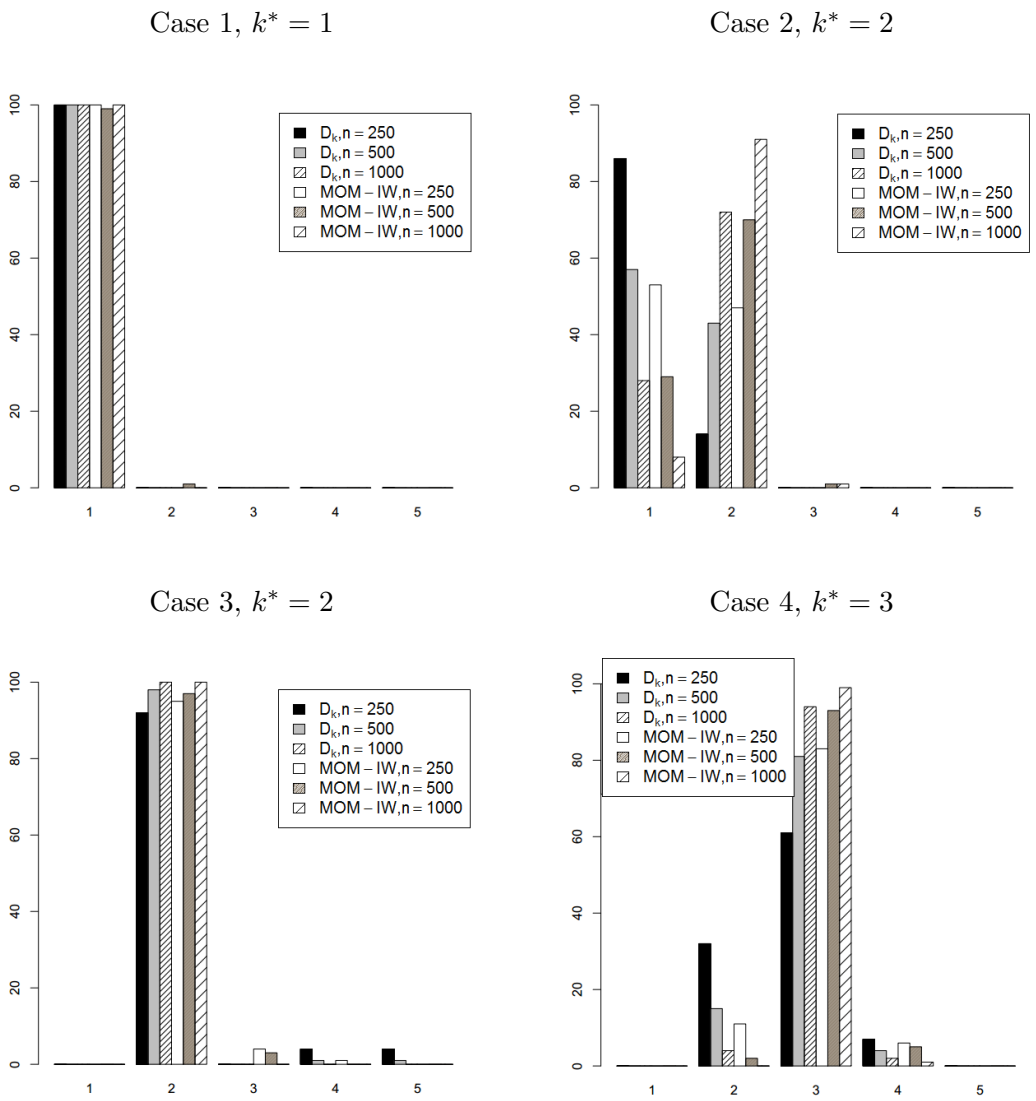


Figure 6.1: Normal mixture. Frequencies of $\hat{k}$ for Cases 1 to 4 (Section 5.1) for 100 data sets, sample sizes of $n = 250$, $n = 500$ and $n = 1000$ using (6.1.4) and MOM-IW priors.

Figure 6.1 illustrates how when n grows D_k was able to select the true number of Normal mixture components even for Cases 2 and 4 where the components are poorly separated.

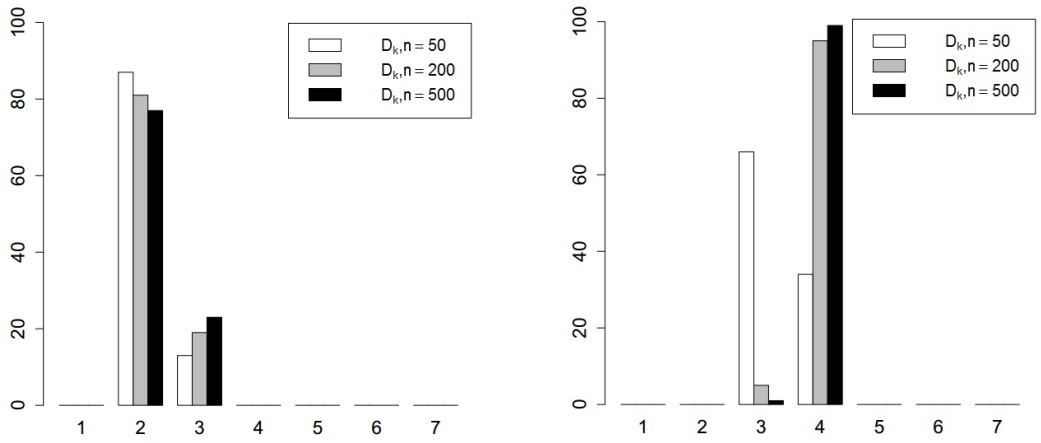


Figure 6.2: Binomial mixture. Frequencies of $\hat{k}$ for 100 data sets. Left: the data considered in Section 5.3 with $k^* = 4$. Right: simulated data from $n = 50$, 200 and 500, $L_{if} = 30$, $\eta_j = 1/4$, $\theta_j = \{0.05, 0.35, 0.65, 0.95\}$ and $k^* = 4$.

Figure 6.1 also shows that although D_k performed well by using the proposed method in Chapter 4 we selected more frequently the true generating model. Figure 6.2 shows the Binomial examples. For the example considered in Section 5.3, D_k chose 3 components thus underfitting the number of components. However, for more separated components D_k was able to select the true generating model when the sample sizes increase.

6.2 Bayes factors for mixtures from cluster occupancies

Consider $\mathbf{z} = \{z_1, ..., z_n\}$ the latent clusters, $n_j = \sum_{i=1}^{n} \mathrm{I}(z_i = j)$ be the number of individuals in cluster j, and $m = \sum_{j=1}^{k} \mathrm{I}(n_j > 0)$ the number of non-empty clusters. We now outline our Empty Cluster Probability (ECP) algorithm, which relies on Proposition 2 below expressing Bayes factors as a ratio of posterior to prior empty cluster probabilities. The result applies to any mixture and prior satisfying the minimal conditions C1-C4 below. In the remainder of this section $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)$ denotes an arbitrary prior for which one wants to obtain posterior model probabilities, e.g. in our examples this is the local prior $\tilde{p}(\boldsymbol{\vartheta}_k \mid M_k)$ and then non-local posterior probabilities are obtained from (4.1.1).

C1 Conditional independence. $p(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \prod_{j=1}^{k} \prod_{z_i=j} p(\mathbf{y}_i \mid \boldsymbol{\theta}_j, \mathcal{M}_k)$

C2 Invariance to label permutations. $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k) = p(\mathbf{y} \mid \psi(\boldsymbol{\vartheta}_k))$ and $p(\mathbf{z} \mid \mathcal{M}_k) = p(\varrho(\mathbf{z}) \mid \mathcal{M}_k)$ for any permutation of component parameters ψ and component indexes ϱ.

C3 Coherence of prior on cluster allocations. $p(\mathbf{z} \mid n_k = 0, \mathcal{M}_k) = p(\mathbf{z} \mid \mathcal{M}_{k-1})$

C4 Coherence of prior on parameters. For any $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \eta_1, \ldots, \eta_{k-1}$, it holds that

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \eta_1, \ldots, \eta_{k-1} \mid \mathbf{z}, \mathcal{M}_{k-1}) = \int p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \eta_1, \ldots, \eta_k \mid \mathbf{z}, \mathcal{M}_k) d\boldsymbol{\theta}_k d\eta_k$$

Conditions C1-C2 hold for the vast majority of mixtures, including mixtures of regressions and most hidden Markov models. Conditions C3-C4 hold for most common priors. For instance C3 holds when $p(\boldsymbol{\eta} \mid \mathcal{M}_k)$ and $p(\boldsymbol{\eta} \mid \mathcal{M}_{k-1})$ are both symmetric $\mathrm{Dir}(q)$ distributions and C4 is satisfied by priors that factor across components, e.g. $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta}; q) \prod_{j=1}^{k} p(\boldsymbol{\theta}_j \mid \mathcal{M}_k)$.

Proposition 2 *Suppose that C1-C4 hold. Then the Bayes factor*

$$B_{k-1,k}(\mathbf{y}) = \frac{\sum_{j=1}^{k} P(n_j = 0 \mid \mathbf{y}, \mathcal{M}_k)/k}{P(n_k = 0 \mid \mathcal{M}_k)}. \tag{6.2.1}$$

Proof. See Appendix A, Section A.7.

Once $B_{k-1,k}(\mathbf{y})$ for $k \in \{2, \ldots, K\}$ are available then $P(\mathcal{M}_k \mid \mathbf{y})$ are obtained as usual. Proposition 2 is easy to implement, e.g. if $p(\boldsymbol{\eta} \mid M_j) = \mathrm{Dir}(\boldsymbol{\eta}; q)$ for all j then

$$a_k = P(n_k = 0 \mid \mathcal{M}_k) = \frac{\Gamma(kq)\Gamma(n + (k-1)q)}{\Gamma((k-1)q)\Gamma(n + kq)}.$$

Further, given draws $\boldsymbol{\vartheta}_k^{(t)} \sim p(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)$ one can obtain Rao-Blackwellised estimates

$$\hat{P}(n_j = 0 \mid \mathbf{y}, \mathcal{M}_k) = \frac{1}{T} \sum_{t=1}^{T} P(n_j = 0 \mid \mathbf{y}, \boldsymbol{\vartheta}_k^{(t)}, \mathcal{M}_k) = \frac{1}{T} \sum_{t=1}^{T} \prod_{i=1}^{n} P(z_i \neq j \mid \mathbf{y}, \boldsymbol{\vartheta}_k^{(t)}, \mathcal{M}_k).$$

and therefore, the estimator of (6.2.1) is given by

$$\tilde{B}_{k-1,k} = \frac{1}{kT} \sum_{j=1}^{k} \sum_{t=1}^{T} \frac{1}{a_k} \prod_{i=1}^{n} P(z_i \neq j | \mathbf{y}, \boldsymbol{\vartheta}_k^{(t)}, \mathcal{M}_k). \tag{6.2.2}$$

Note also that ECP only requires cluster probabilities, hence it remains valid for non-conjugate models. We used the estimator in (4.1.2)-(4.1.4) for many of our examples, as it was used in an earlier version of this manuscript, but we found that the ECP estimator was highly efficient (Section 6.2.2). To compute the Bayes factor comparing $k-1$ versus k components under NLPs we use the estimator suggested in 4.1.1 and 6.2.2 under LPs as follows

$$\hat{B}_{k-1,k} = \tilde{B}_{k-1,k} \frac{\sum_{t=1}^{T} \omega(\boldsymbol{\vartheta}_{k-1}^{(t)})}{\sum_{t=1}^{T} \omega(\boldsymbol{\vartheta}_{k}^{(t)})}. \tag{6.2.3}$$

Proposition 2 is of independent interest to help discard unoccupied clusters in overfitted mixtures. It suggests that the threshold on posterior empty cluster probabilities should depend on the corresponding prior empty cluster probabilities. The latter are a function of n, k and q, hence using fixed thresholds may be suboptimal. Note also that Proposition 2 can be used to compare structurally different models. For instance let B_{k1} be the Bayes factor between a k-component unequal-covariance Normal mixture vs. a one-component Normal, and B_{k1}^{c} that for a k-component common-covariance Normal mixture vs. a one-component Normal. Then B_{k1}/B_{k1}^{c} is the Bayes factor comparing k components with unequal vs. equal covariances. Similarly one could combine the Bayes factor between a one-component Normal vs. a one-component T (which is easy to compute) with Proposition 2 to obtain Bayes factors between any k-component Normal vs. T mixture. That is, the ECP estimator is connected to empty cluster probabilities but really is a tool to obtain $P(\mathcal{M}_k \mid \mathbf{y})$ and hence remains applicable in more general settings.

6.2.1 Comparison with other alternatives

We simulated a single data set of $n = 200$ observations from Cases 1 and 3 in Section 5.1 and computed 50 times $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir priors using the ECP estimator and the Marin and Robert (2008) estimator given by (4.1.4) in Section 4.1. Figures 6.3-6.4 shows how the medians of the ECP estimator and the Marin and Robert (2008) estimator with $k = \{1, ..., 4\}$ are similar however the ECP estimator instead of the Marin and Robert (2008) estimator produces higher precision estimates. To compute $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using the ECP estimator we implement the bfnormmix function given in the R package `mombf` (Rossell et al., 2018).

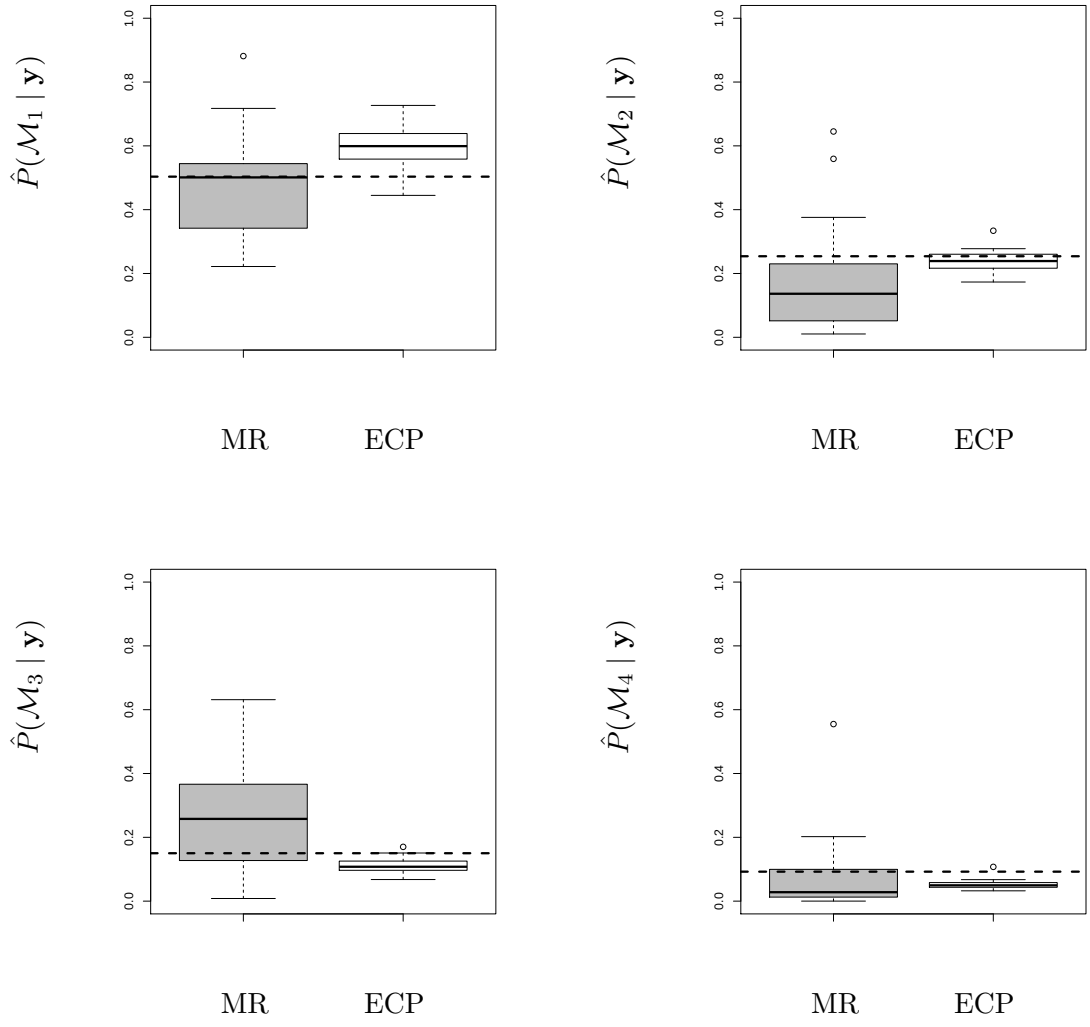


Figure 6.3: Boxplots display 50 independent estimates based on separate MCMC runs ($T = 10,000$ iterations after a $T/10$ burn-in each). Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir using the Marin and Robert (2008) (MR) estimator (gray) and ECP estimator (white) for $n = 200$ observations in simulation Case 1, $k^* = 1$. Dashed line indicate $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir obtained by simulating $1,000,000$ values from the prior and averaging the likelihood.

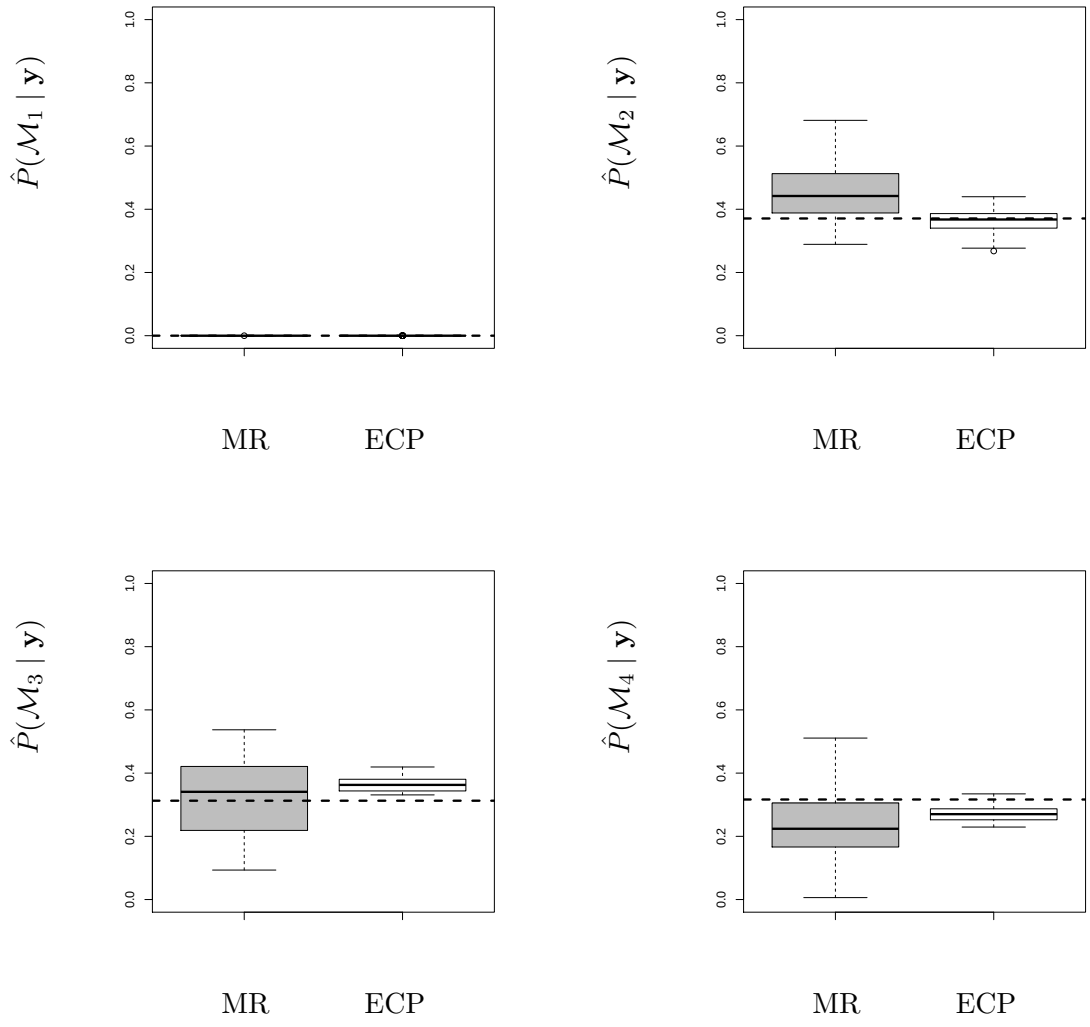


Figure 6.4: Boxplots display 50 independent estimates based on separate MCMC runs ($T = 10,000$ iterations after a $T/10$ burn-in each). Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir using the Marin and Robert (2008) estimator (gray) and ECP estimator (white) for $n = 200$ observations in simulation Case 3, $k^* = 2$. Dashed line indicate $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ under Normal-IW-Dir obtained by simulating $1,000,000$ values from the prior and averaging the likelihood.

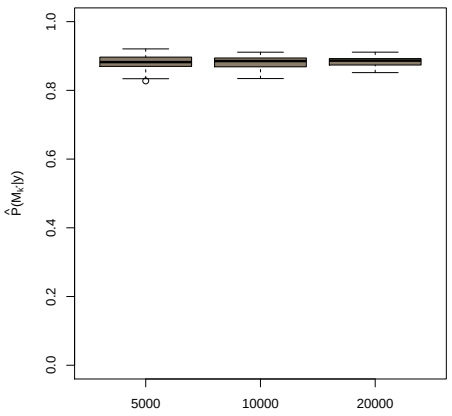
6.2.2 Computational cost and precision across MCMC runs for the ECP estimator

We run 50 simulations under Cases 1, 3, 5 and 7 for $n \in \{200, 1000\}$, for each dataset we obtained $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, 2, 3\}$ using the ECP estimator (9,000 iterations after a 1,000 burn-in) both for MOM-IW-Dir and Normal-IW-Dir with unequal covariances. Table 6.1 reports the average posterior probabilities and median run time on a laptop running OS X 10.11.6 with 1.6 GHz processor and 8Gb 1600MHz DDR3. The runtime in Table 6.1 corresponds to the total time of obtaining $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ for all k and both priors, using function bfnormmix in R package `mombf` (Rossell et al., 2018).

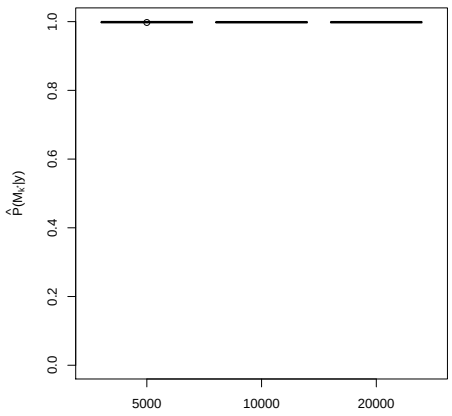
		MOM-IW-Dir			Normal-IW-Dir			
	n	k=1	k=2	k=3	k=1	k=2	k=3	CPU time
Case 1, $k^* = 1$	200	0.860	0.061	0.079	0.701	0.190	0.109	1.8 sec.
	1000	0.989	0.010	0.001	0.893	0.089	0.018	8.3 sec.
Case 3, $k^* = 2$	200	0.000	0.727	0.273	0.000	0.592	0.408	1.8 sec.
	1000	0.000	0.933	0.067	0.000	0.776	0.224	8.4 sec.
Case 5, $k^* = 1$	200	0.937	0.060	0.003	0.871	0.110	0.019	2.7 sec.
	1000	0.994	0.006	0.000	0.925	0.070	0.006	12.9 sec.
Case 7, $k^* = 2$	200	0.611	0.343	0.046	0.675	0.277	0.048	2.7 sec.
	1000	0.000	0.955	0.045	0.000	0.879	0.121	13.3 sec.

Table 6.1: Simulation study. Mean $P(\mathcal{M}_k \mid \mathbf{y})$ for $k \in \{1, 2, 3\}$ and Cases 1, 3, 5 and 7 under MOM-IW-Dir and Normal-IW-Dir priors. Median CPU time (seconds) to compute $P(\mathcal{M}_k \mid \mathbf{y})$ for both priors and all k with bfnormmix in R package `mombf` (Rossell et al., 2018).

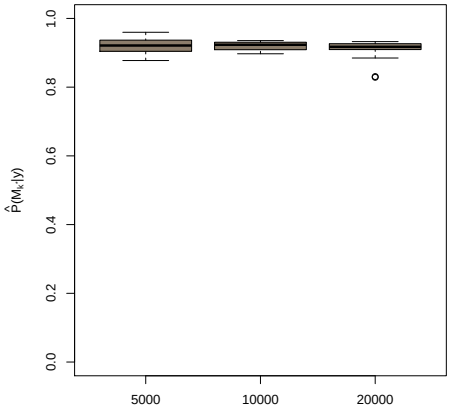
To assess the numerical precision of our estimates we simulated a single dataset under each scenario and obtained 20 estimates $\hat{P}(\mathcal{M}_{k^*} \mid \mathbf{y})$ and $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ from independent MCMC runs. In Figures 6.5-6.6 we consider $T \in \{5000, 10000, 20000\}$ after a T/10 burn-in each. The Figures 6.5-6.6 illustrate how $T = 10,000$ are enough to obtain $\hat{P}(\mathcal{M}_{k^*} \mid \mathbf{y})$ with high precision (except in Case 7 with $n = 200$ where even $T = 20,000$ seems to be in adequate). In Figures 6.7-6.8 and 6.9-6.10 we consider $T = 10,000$ and $T = 20,000$ after a T/10 burn-in each where $k \in \{1, 2, 3\}$. Figures 6.7-6.10 show how the precision for $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ was very high in all settings, particularly for $n = 1000$.



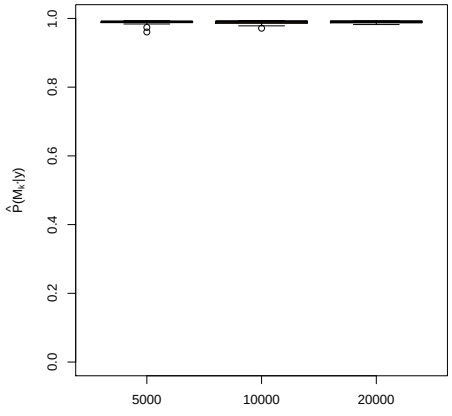
(a) Case 1, $k^* = 1$, $n = 200$



(b) Case 1, $k^* = 1$, $n = 1000$

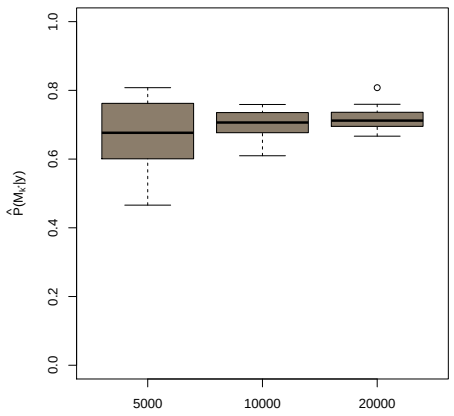


(c) Case 3, $k^* = 2$, $n = 200$

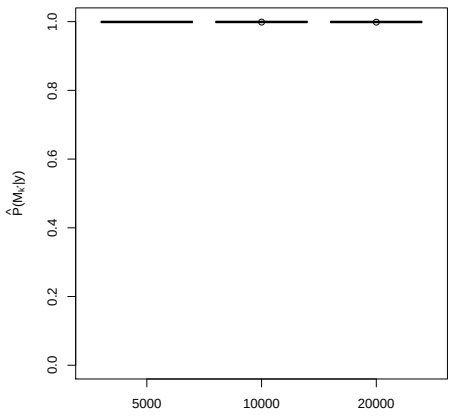


(d) Case 3, $k^* = 2$, $n = 1000$

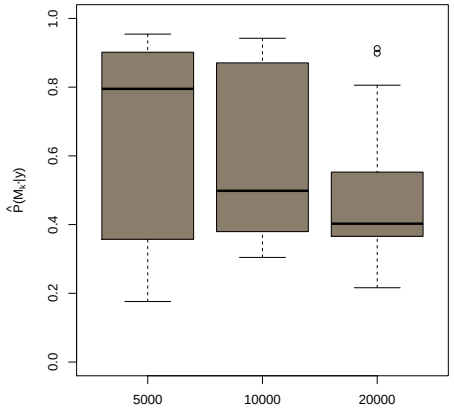
Figure 6.5: Precision of $\hat{P}(\mathcal{M}_{k^*} \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. Boxplots display 20 independent estimates based on separate MCMC runs ($T \in \{5000, 10000, 20000\}$ iterations after a $T/10$ burn-in each).



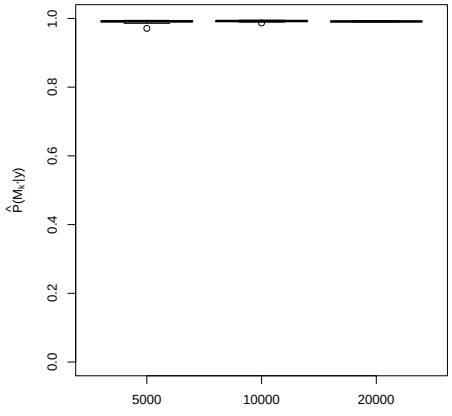
(a) Case 5, $k^* = 1$, $n = 200$



(b) Case 5, $k^* = 1$, $n = 1000$

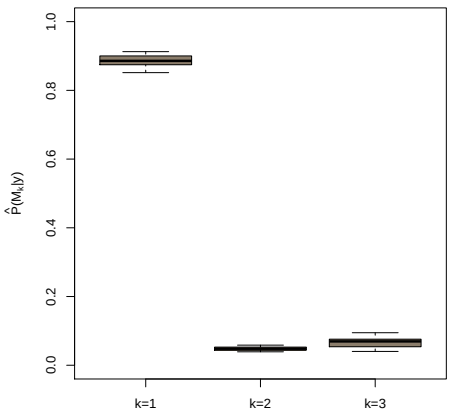


(c) Case 7, $k^* = 2$, $n = 200$



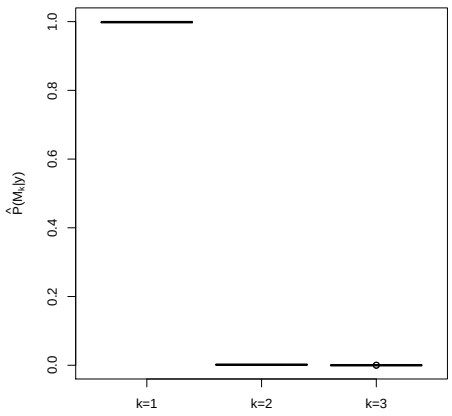
(d) Case 7, $k^* = 2$, $n = 1000$

Figure 6.6: Precision of $\hat{P}(\mathcal{M}_k^* \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. Boxplots display 20 independent estimates based on separate MCMC runs ($T \in \{5000, 10000, 20000\}$ iterations after a $T/10$ burn-in each).



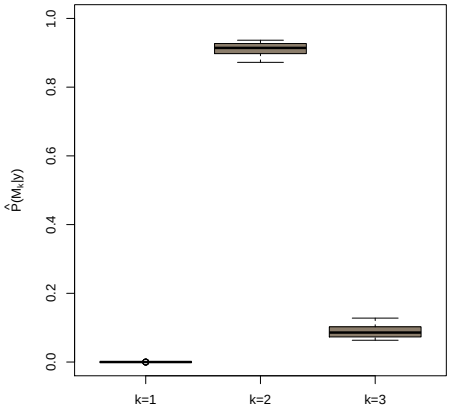
T

(a) Case 1, $k^* = 1$, $n = 200$



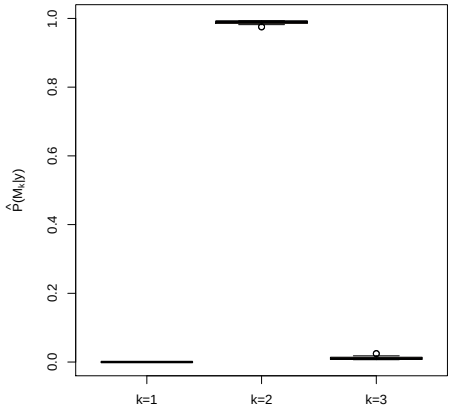
T

(b) Case 1, $k^* = 1$, $n = 1000$



T

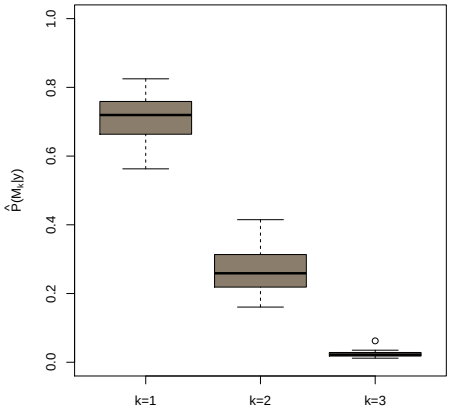
(c) Case 3, $k^* = 2$, $n = 200$



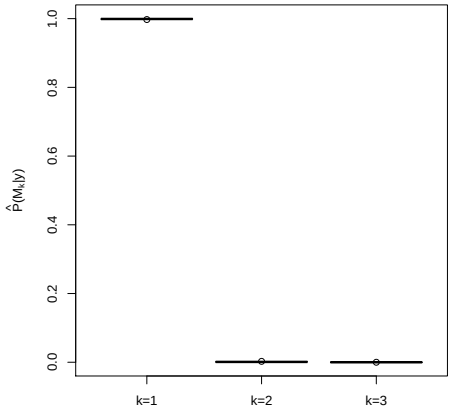
T

(d) Case 3, $k^* = 2$, $n = 1000$

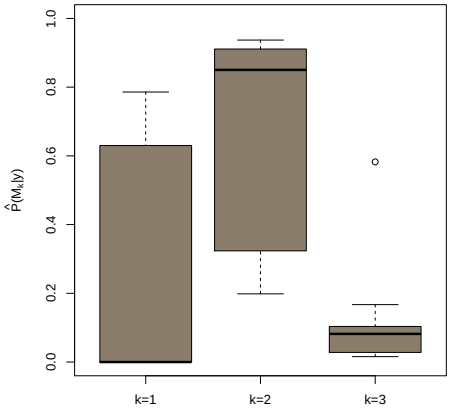
Figure 6.7: Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. Boxplots display 20 independent estimates based on separate MCMC runs ($T = 10,000$ iterations after a $T/10$ burn-in each).



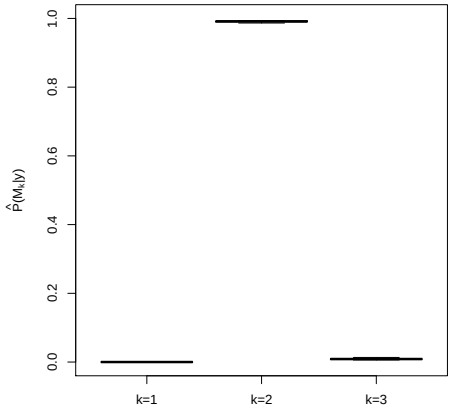
(a) Case 5, $k^* = 1$, $n = 200$



(b) Case 5, $k^* = 1$, $n = 1000$

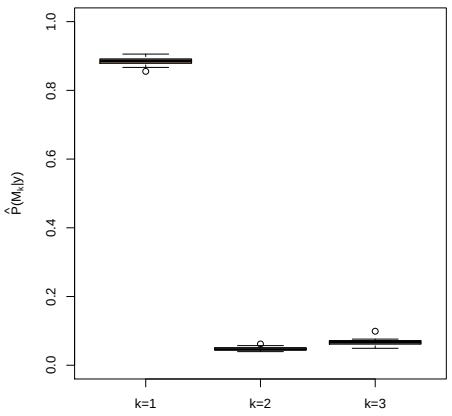


(c) Case 7, $k^* = 2$, $n = 200$



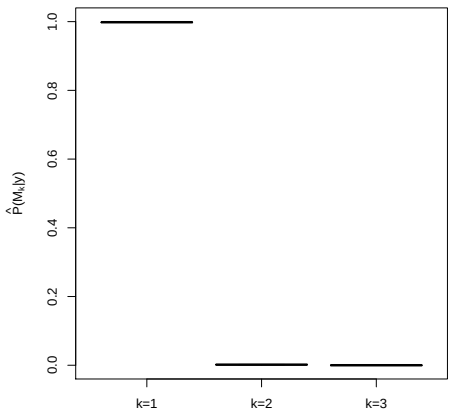
(d) Case 7, $k^* = 2$, $n = 1000$

Figure 6.8: Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. Boxplots display 20 independent estimates based on separate MCMC runs ($T = 10,000$ iterations after a $T/10$ burn-in each).



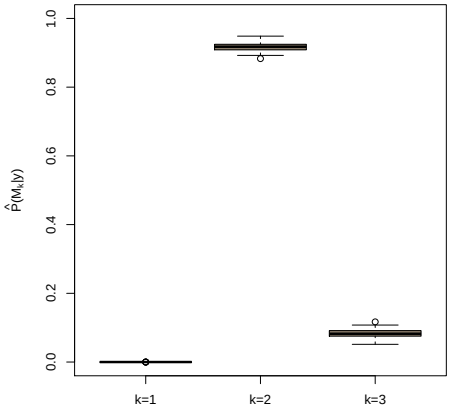
T

(a) Case 1, $k^* = 1$, $n = 200$



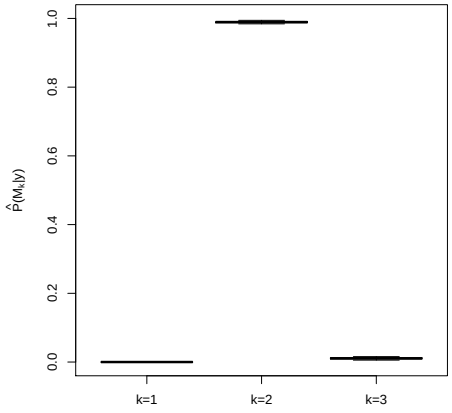
T

(b) Case 1, $k^* = 1$, $n = 1000$



T

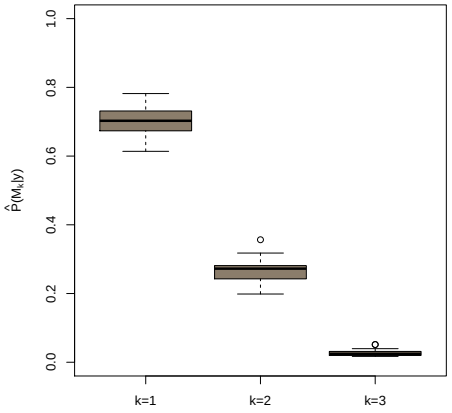
(c) Case 3, $k^* = 2$, $n = 200$



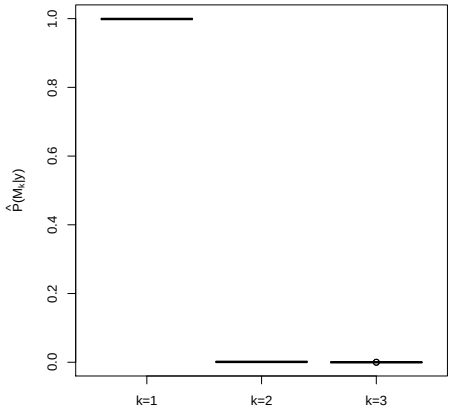
T

(d) Case 3, $k^* = 2$, $n = 1000$

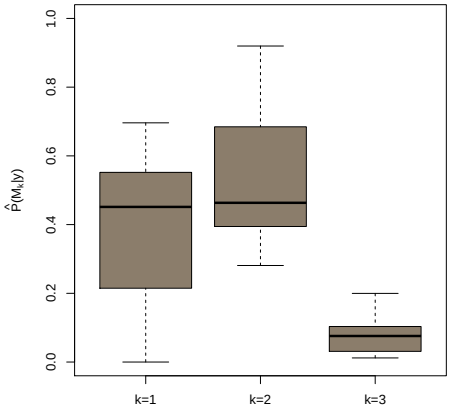
Figure 6.9: Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 1 and 3. Boxplots display 20 independent estimates based on separate MCMC runs ($T = 20,000$ iterations after a $T/10$ burn-in each).



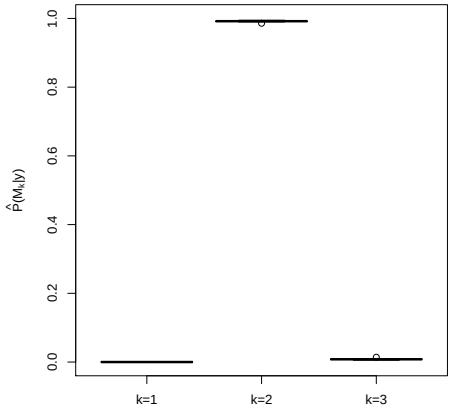
(a) Case 5, $k^* = 1$, $n = 200$



(b) Case 5, $k^* = 1$, $n = 1000$



(c) Case 7, $k^* = 2$, $n = 200$



(d) Case 7, $k^* = 2$, $n = 1000$

Figure 6.10: Precision of $\hat{P}(\mathcal{M}_k \mid \mathbf{y})$ using ECP estimator in simulation Cases 5 and 7. Boxplots display 20 independent estimates based on separate MCMC runs ($T = 20,000$ iterations after a $T/10$ burn-in each).

Chapter 7

Applications

In this chapter we compare the performance of our MOM-IW-Dir and MOM-Beta-Dir priors with respect to Normal-IW-Dir and Beta-Dir, BIC, AIC, sBIC and overfitted mixtures and repulsive overfitted mixtures using textbook and real applications. Section 7.1 presents the Old-Faithful dataset. In Sections 7.2-7.3 we analyze a flow cytometry experiment and Fisher's Iris data for which there is a known ground truth. In Section 7.4 we offer a comparison with overfitted and repulsive overfitted mixtures. Finally, in Section 7.5 we analyze a USA political blog dataset via product Binomial mixtures. We assessed MCMC convergence for all parameters after a burn-in period via MCMC iteration plots (see Appendix B).

7.1 Old Faithful

We briefly describe this classical example to illustrate potential issues with poorly-separated components. The results are in Table 7.1 and Figures 7.2-7.3.

Figure 7.1: Old Faithful: the biggest cone-type geyser located in the Yellowstone National Park, Wyoming, United States.

The Old Faithful is a cone-type geyser in the Yellowstone National Park (Figure 7.1). We seek clusters in a dataset with $n = 272$ eruptions recording their duration and the time to the next eruption (dataset faithful in R). We considered up to $K = 6$ Normal components either with equal or unequal covariance matrices.

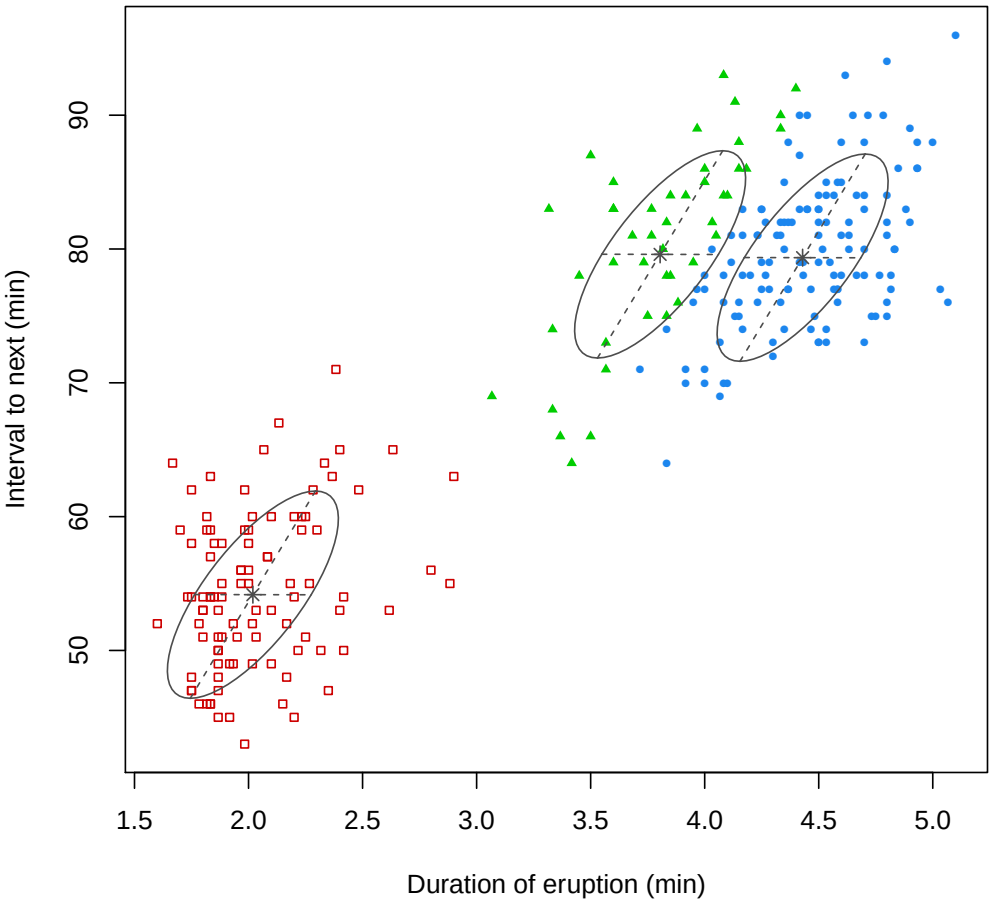


Figure 7.2: Classification and contours for the model chosen by MOM-IW-Dir for Faithful data set.

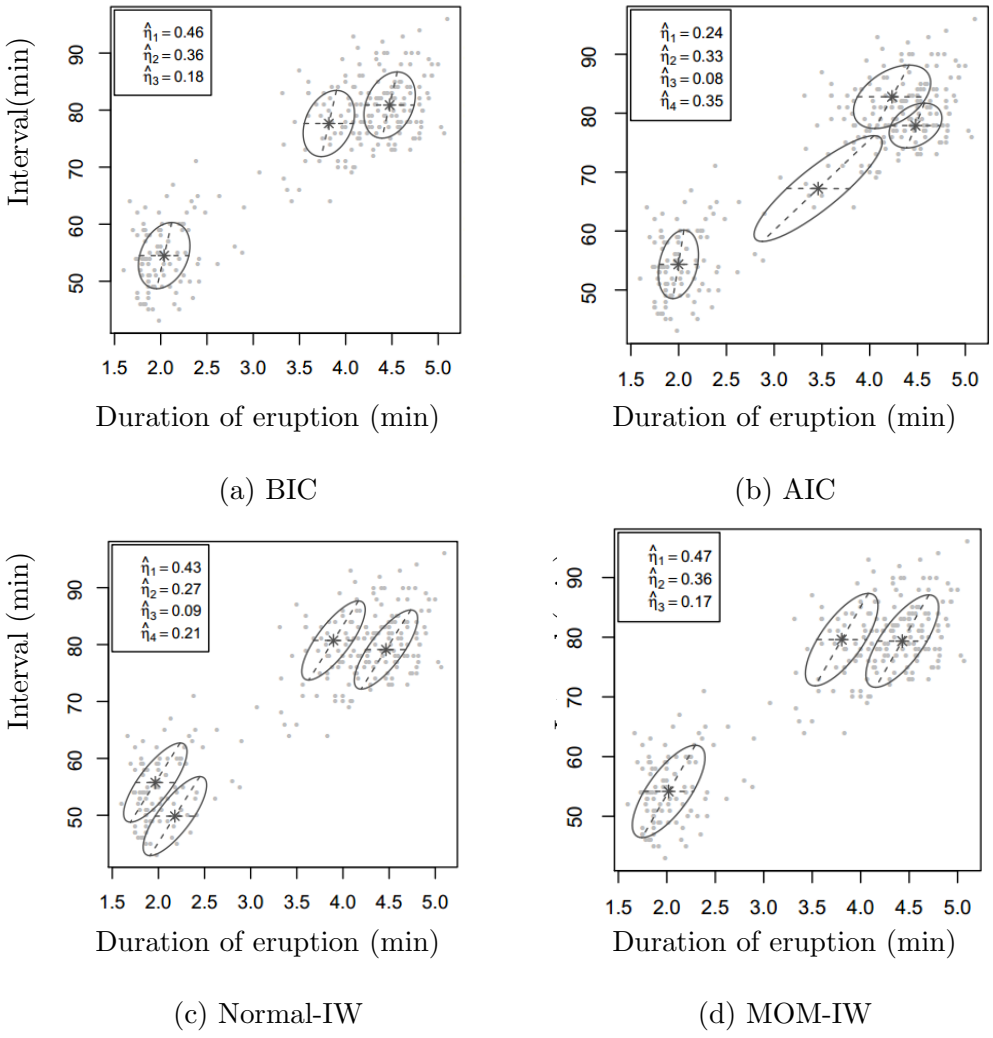


Figure 7.3: Faithful dataset. Contours for the model chosen by (a) BIC and (b) AIC (top right), (c) Normal-IW (bottom left) and (d) MOM-IW (bottom right). Points indicate the data.

Table 7.1 shows how the Our MOM-IW-Dir selected $k = 3$ equal-covariance components with 0.967 posterior probability (Figure 7.2). The Normal-IW-Dir chose $k = 4$ with 0.473 posterior probability, this resulted from splitting an MOM-IW-Dir component in the lower-left corner into two. The sBIC and BIC chose $\hat{k} = 3$ components with roughly the similar location as the MOM-IW-Dir, though their shapes were slightly different, whereas AIC returned $\hat{k} = 4$ (Figure 7.3).

Table 7.1: Faithful dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under Normal-IW-Dir, MOM-IW-Dir, BIC, AIC and sBIC under $\Sigma_i \neq \Sigma_j$.

		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
$\Sigma_j = \Sigma$	1	0.000	0.000	-558.006	-548.992	
	2	0.000	0.000	-416.805	-402.382	
	3	0.132	**0.967**	**-411.356**	-391.524	
	4	**0.473**	0.000	-419.748	-394.507	
	5	0.353	0.000	-418.019	-387.369	
	6	0.042	0.000	-427.821	-391.763	
$\Sigma_i \neq \Sigma_j$	2	0.000	0.000	-415.291	-395.459	-419.103
	3	0.000	0.000	-422.609	-391.960	**-415.938**
	4	0.000	0.000	-425.370	**-383.903**	-417.278
	5	0.000	0.000	-439.754	-387.470	-420.569
	6	0.000	0.000	-448.896	-385.795	-422.231

7.2 Cytometry data

We analysed the Graf-versus-Host flow cytometry data in Brinkman et al. (2007), an experiment used for cell counting, *e.g.* to diagnose diseases. The data contain $p = 4$ variables called CD3, CD4, CD8b and CD8 (Figure 7.4). The study goal was to find cell subpopulations with positive CD3, CD4 and CD8b (CD3+/CD4+/CD8b+), i.e. high values in the first three variables. Interestingly, the authors created a control sample designed not to contain any CD4+/CD8b+ cells.

Following the analysis in Baudry et al. (2012), we selected the $n = 1,126$ cells in the control sample for which CD3 > 280.

Figure 7.5 plots (CD4,CD8b) values and the solution chosen by BIC, AIC and Normal-IW-Dir, MOM-IW-Dir. The first three methods identified a CD4+ and CD8b+ subpopulation that as discussed is not there by design, whereas it was not present in the MOM-IW-Dir solution. In contrast, sBIC supported six components with unequal covariances. Intuitively, the spurious CD4+ and CD8b+ cluster contains a few outlying observations, and our MOM-IW-Dir penalises such a low-weight component. This is an interesting contrast to our other examples where the issue was having poorly-separated, rather than low-weight, components. These results illustrate the benefits of jointly penalising small weights and overlapping components. See Table 7.2 for further details, e.g. the Normal-IW-Dir and MOM-IW-Dir chose $k = 3$ with 0.928 and 0.995 posterior probability, respectively.

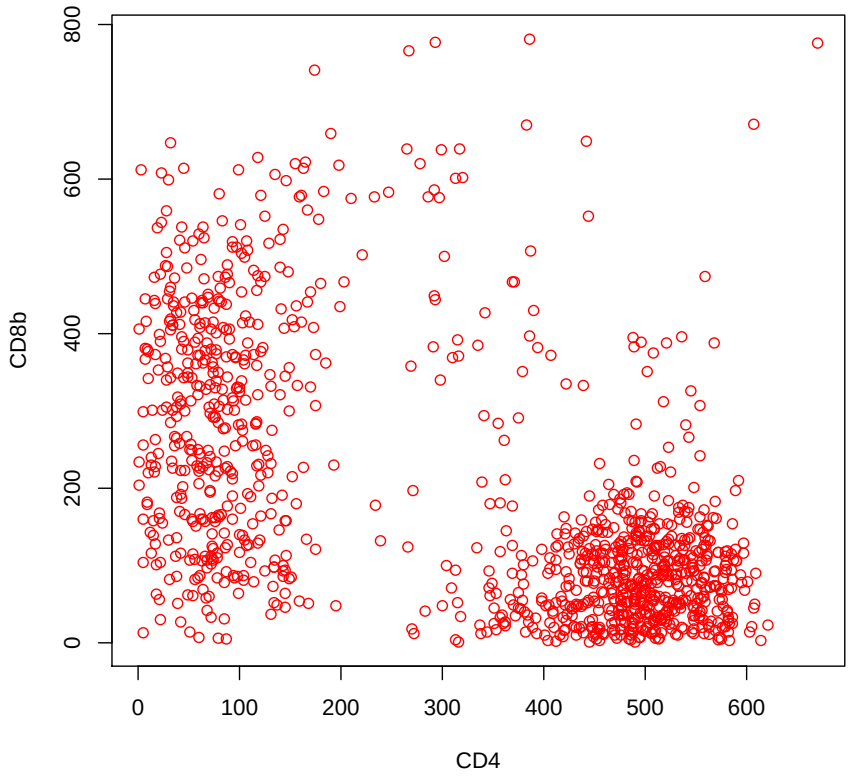


Figure 7.4: Cytometry data-set with the variables CD4 and CD8b.

Table 7.2: Cytometry dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under Normal-IW-Dir, MOM-IW-Dir, BIC and BIC and sBIC under $\Sigma_i \neq \Sigma_j$.

		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
$\Sigma_j = \Sigma$	1	0.000	0.000	-28337.23	-28295.02	
	2	0.000	0.000	-27720.64	-27665.86	
	3	0.000	0.000	-27541.73	-27474.39	
	4	0.000	0.000	-27443.22	-27363.31	
	5	0.000	0.000	-27271.67	-27179.19	
	6	0.000	0.000	-27226.41	-27121.36	
$\Sigma_i \neq \Sigma_j$	2	0.072	0.005	-27357.56	-27277.65	-37869.06
	3	**0.928**	**0.995**	**-27015.35**	-26897.74	-36478.20
	4	0.000	0.000	-27048.60	-26893.29	-35247.11
	5	0.000	0.000	-27041.50	-26848.50	-34415.96
	6	0.000	0.000	-27075.18	**-26844.48**	**-33888.20**

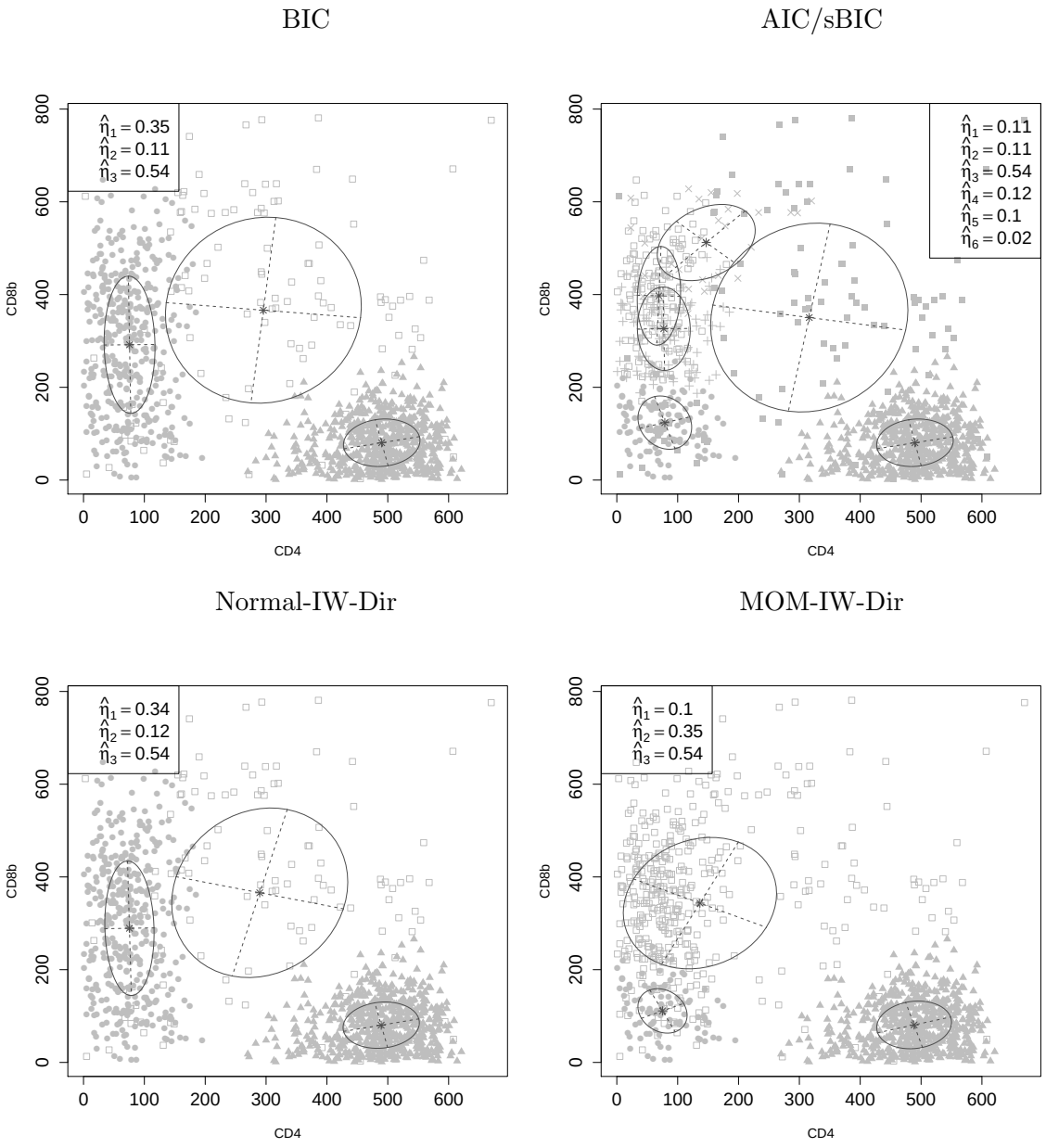


Figure 7.5: Projection of the variables CD4 and CD8b for the Cytometry data-set, classification of observations and contours using EM algorithm for BIC and AIC (top), and under Normal-IW-Dir and MOM-IW-Dir (bottom)

7.3 Fisher's Iris data

We present another classical dataset by Fisher (1936) for the practical reason that there is a ground truth for the underlying number of subpopulations. The data contain four variables ($p = 4$) measuring the dimensions of $n = 150$ iris flowers. The plants are known to belong to $k^* = 3$ species, setosa, versicolor and virginica, each with 50 observations (Figure 7.6). We compare the ability of the various methods to recover these three species in an unsupervised fashion. We considered up to $K = 6$ Normal components with either equal or unequal covariances.

Table 7.3 provides posterior model probabilities. The BIC and sBIC supported $\hat{k} = 2$ and $\hat{k} = 4$ components with unequal covariances, respectively. Upon inspection the BIC solution merged the versicolor and virginica species into a single component, akin to its lack of sensitivity observed in Section 5.1, whereas the sBIC split the versicolor specie into two components. The AIC supported $\hat{k} = 6$ with unequal covariances.

Table 7.3: Iris dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for 11 models with $k \in \{1, \ldots, 6\}$ and either homogeneous ($\Sigma_j = \Sigma$) or heterogeneous ($\Sigma_i \neq \Sigma_j$) under Normal-IW-Dir, MOM-IW-Dir, BIC and BIC and sBIC under $\Sigma_i \neq \Sigma_j$.

		Normal-IW-Dir	MOM-IW-Dir	BIC	AIC	sBIC
	k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$			
	1	0.000	0.000	-414.989	-393.915	
$\Sigma_j = \Sigma$	2	0.000	0.000	-344.049	-315.448	
	3	**0.809**	**1.000**	-316.483	-280.355	
	4	0.029	0.000	-295.705	-252.051	
	5	0.132	0.000	-302.465	-251.284	
	6	0.030	0.000	-310.909	-252.201	
$\Sigma_i \neq \Sigma_j$	2	0.000	0.000	**-287.009**	-243.355	-415.449
	3	0.000	0.000	-290.420	-224.186	-410.122
	4	0.000	0.000	-314.483	-225.669	**-408.839**
	5	0.000	0.000	-341.910	-230.517	-414.190
	6	0.000	0.000	-355.786	**-221.813**	-422.209

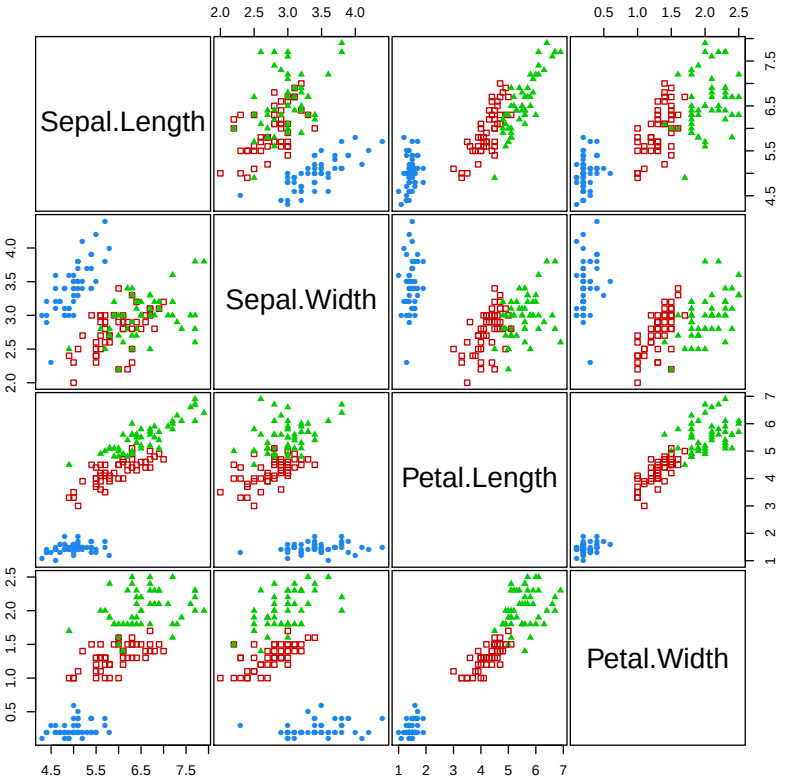


Figure 7.6: Top: The species, setosa, versicolor and virginica in the Fisher's Iris data set. Bottom: Classification for the model chosen by MOM-IW-Dir for Iris data set.

The AIC supported $\hat{k} = 6$ with unequal covariances. Both the Normal-IW-Dir and our MOM-IW-Dir chose $\hat{k} = 3$ (see Figure 7.6), albeit the evidence under the former was weaker ($P^L(\mathcal{M}_3 \mid \mathbf{y}) = 0.81$ and $P(\mathcal{M}_3 \mid \mathbf{y}) = 1$ respectively).

7.4 Comparison with overfitted and repulsive overfitted mixtures

Table 7.4 and Table 7.5 summarise the results from analysing the misspecified student-T mixture in Section 5.2 and the datasets from Sections 7.1-7.3 with overfitted mixtures and repulsive overfitted mixtures (respectively) with $\Sigma_j = \Sigma$ as in Petralia et al. (2012). Here repulsion was induced by a pMOM penalty where g is set to its default in Section 3.2. We set $k = 6$ and report the posterior distribution of the number of empty components (with no assigned observations) from the MCMC output. Note that $k = 6$ favors overfitted mixtures as our analyses in Sections 5.2 and 7.1-7.3 suggested less than 6 components. To assess sensitivity we tested prior parameter values $q = 1$ (no shrinkage), $q = 0.01$ (satisfying Rousseau and Mengersen (2011) and Gelman et al. (2013)) and 3×10^{-8} (proposed by Havre et al. (2015)). We observed little differences between overfitted and repulsive overfitted mixtures. As expected in general smaller q led to less occupied components in the posterior, except in the cytometry data where the posterior focused on 6 components for all q. Note that $q = 3 \times 10^{-8}$ recovered the true $k^* = 3$ in the misspecified example from Section 5.2, whereas this was not the case in the Iris and Cytometry data that truly contain 3 subpopulations.

Table 7.4: Posterior for the distribution on non-empty components m in overfitted mixtures under $\Sigma_j = \Sigma$, the misspecified student-T mixture, Faithful, Iris and Cytometry data.

	$\hat{P}(m \mid \mathbf{y}, \mathcal{M}_6)$					
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
			$q = 1$			
Misspecified	0.00	0.00	0.00	0.00	0.07	0.93
Faithful	0.00	0.00	0.00	0.01	0.15	0.85
Fisher's Iris	0.00	0.99	0.01	0.00	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00
			$q = 0.01$			
Misspecified	0.00	0.00	0.03	0.35	0.56	0.06
Faithful	0.00	0.00	0.63	0.31	0.04	0.02
Fisher's Iris	0.00	1.00	0.00	0.00	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00
			$q = 3.10^{-8}$			
Misspecified	0.00	0.00	0.95	0.00	0.00	0.05
Faithful	0.00	0.00	0.96	0.00	0.01	0.03
Fisher's Iris	0.00	1.00	0.00	0.00	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00

Table 7.5: Posterior distribution on non-empty components m in repulsive overfitted mixtures under $\Sigma_j = \Sigma$. The misspecified student-T mixture, Faithful, Iris and Cytometry.

	$\hat{P}(m \mid \mathbf{y}, \mathcal{M}_6)$					
	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
			$q=1$			
Misspecified	0.00	0.00	0.00	0.00	0.02	0.98
Faithful	0.00	0.00	0.00	0.00	0.26	0.74
Iris	0.00	0.99	0.00	0.01	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00
			$q=0.01$			
Misspecified	0.00	0.00	0.00	0.35	0.63	0.02
Faithful	0.00	0.00	0.76	0.23	0.01	0.00
Iris	0.00	1.00	0.00	0.00	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00
			$q=3.10^{-8}$			
Misspecified	0.00	0.00	0.83	0.00	0.00	0.17
Faithful	0.00	0.00	0.99	0.01	0.00	0.00
Iris	0.00	1.00	0.00	0.00	0.00	0.00
Cytometry	0.00	0.00	0.00	0.00	0.00	1.00

The results for the faithful data matched those of our MOM-IW, but in these data there is not ground truth and it is hence hard to judge the quality of any given answer.

7.5 Political blog data

We illustrate product Binomial mixtures using a dataset on $n = 773$ USA political blogs from 2008 (Chang, 2015). Each blog provides word counts (how many times a given word was used). To facilitate interpretation we combined similar words (e.g. america, american and americans, see Table 7.6 and Figure 7.7) and selected the $p = 234$ words with overall frequency above 100. We fitted the product Binomial mixture $y_{if} \mid z_i = j, \theta_{jf} \sim \text{Bin}(\theta_{jf}, L_i)$, where $L_i = \sum_{f=1}^{p} y_{if}$ is the total number of words in blog $i = 1, ..., 773$. We considered MOM-Beta-Dir and Beta-Dir priors and the BIC and AIC model choice criteria. The MOM-Beta-Dir parameters were set to the default in Section 3.2, obtaining $a = 1/2$, $g = 2.02$ whereas as a local prior we chose the Beta$(1, 1)$ to match the prior variance of the MOM-Beta.

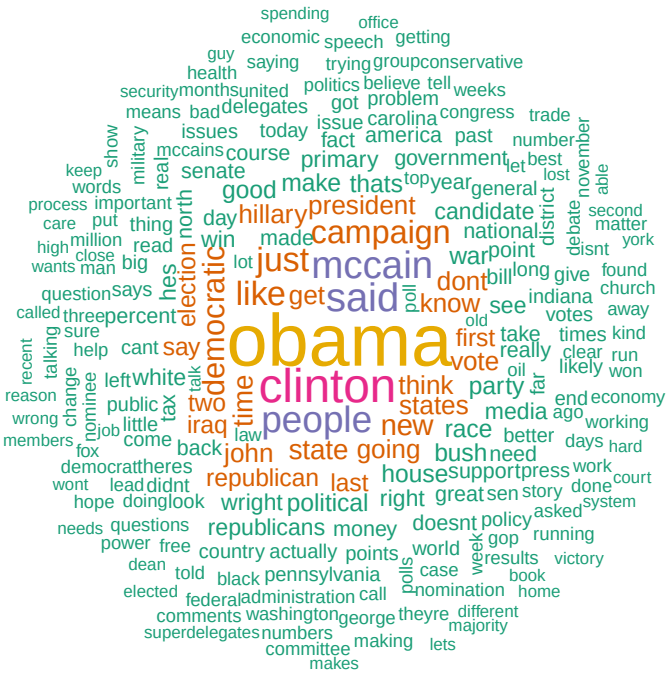


Figure 7.7: Word cloud for the political blog data set where sizes are increasing with frequency of use.

The MOM-Beta-Dir and Beta-Dir chose $\hat{k} = 2$ and $\hat{k} = 4$ both with posterior probability one (up to rounding), whereas BIC and AIC chose $\hat{k} = 3$ and $\hat{k} = 6$ respectively (Table 7.8). To assess the inferred components, these data contain an independent labeling that classifies blogs either as liberal or conservative. Figure 7.8 displays posterior cluster probabilities (Algorithm 4) for liberal and conservative blogs. Interestingly under the MOM-Beta prior conservative blogs fell mostly in Component 1. Figure 7.9 shows the most characteristic words for each MOM-Beta-Dir component (chi-square residual$>$ 2 when cross-tabulating word count versus assigned component). For instance, "war, iraq, tax, government" are representative of Component 1 whereas "polls, votes, percent, delegates" are representative of Component 2. In contrast, under the Beta prior Components 2 and 4 showed a similar distribution for liberal and conservative blogs and similarly for Components 1 and 3 obtained under the BIC. The clusters returned by the AIC did not show appreciable differences between conservative or liberal blogs. We emphasize that there is no ground truth for these data, hence our discussion is simply meant as a qualitative assessment.

Table 7.6: The similar words combined into a single word to facilitate interpretation for USA political blogs dataset.

clinton	clintons	
obama	obamas	barack
america	american	americans
candidate	candidates	
democratic	democrats	
new	news	
president	presidential	
senate	senator	
year	years	
vote	voters	
thing	things	

Table 7.7: 20 most representative words for each component with largest chi-square residual values.

Component 1	Component 2
people	obama
war	clinton
just	vote
said	election
president	democratic
mccain	race
like	percent
iraq	hillary
house	win
new	campaign
government	votes
day	polls
tax	indiana
bush	mccain
know	state
make	primary
america	carolina
time	delegates
get	said
media	results

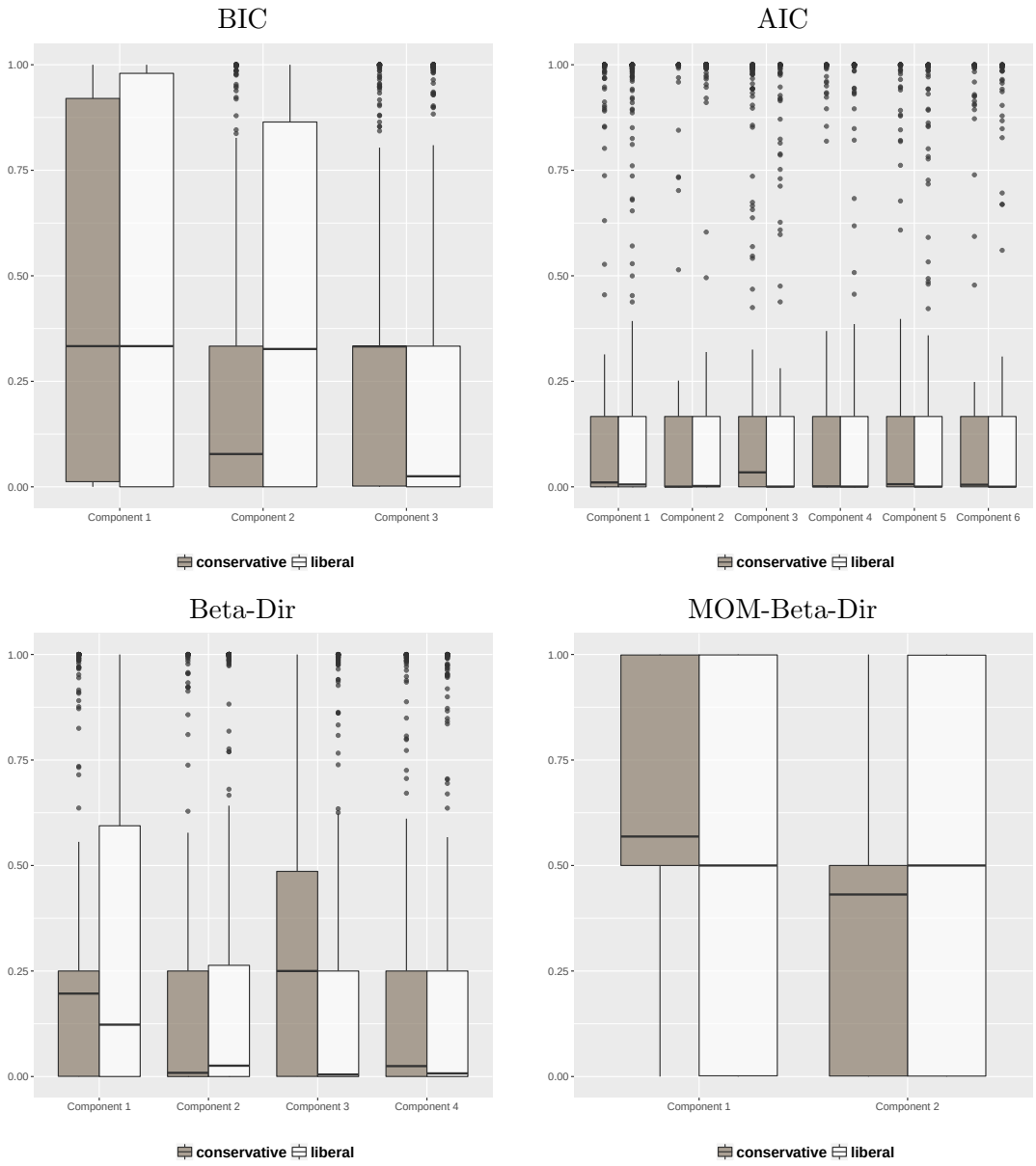


Figure 7.8: Posterior cluster probabilities $p(z_i = j|\mathbf{y}, \mathcal{M}_j)$ under BIC, AIC and Beta-Dir, MOM-Beta-Dir for documents labelled as conservative or liberal

Table 7.8: USA political blogs dataset. $P(\mathcal{M}_k \mid \mathbf{y})$ for 6 models with $k \in \{1, \ldots, 6\}$ under Beta-Dir, MOM-Beta-Dir, BIC and BIC.

	MOM-Beta-Dir	Beta-Dir	BIC	AIC
k	$P(\mathcal{M}_k \mid \mathbf{y})$	$P(\mathcal{M}_k \mid \mathbf{y})$		
1	0.000	0.000	-257405.2	-256317.0
2	**1.000**	0.000	-255488.8	-253307.8
3	0.000	0.000	**-255329.7**	-252055.9
4	0.000	**1.000**	-255358.8	-250992.2
5	0.000	0.000	-255712.2	-250252.7
6	0.000	0.000	-256366.0	**-249813.7**

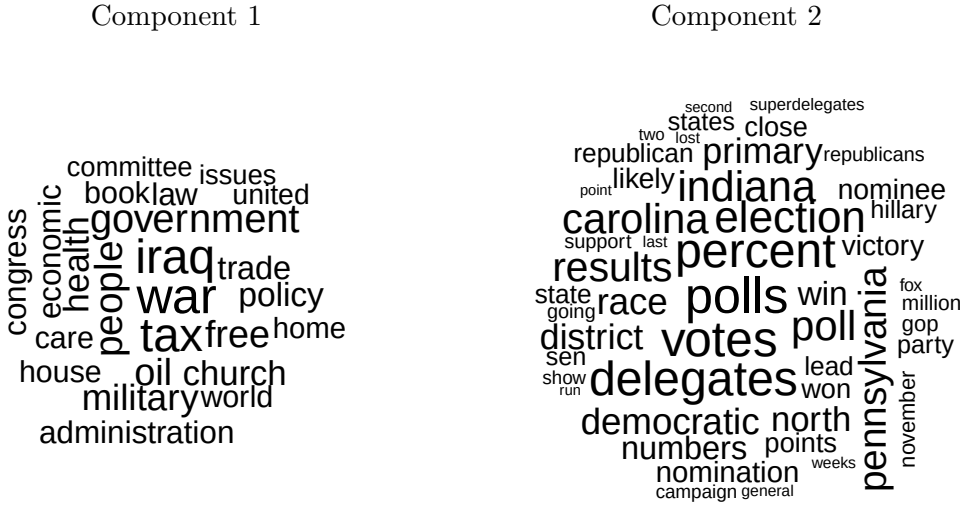


Figure 7.9: Political blog data. Each blog was assigned to its most probable component under a MOM-Beta prior. Word sizes based on chi-square residuals from cross-tabulating word frequency versus assigned component

Chapter 8

Conclusions and future work

According to this thesis, we have the following conclusions:

(i) The primary reasons for using NLPs to select mixture components are encouraging solutions that balance parsimony and sensitivity, and also facilitate interpretation in terms of well-separated subpopulations. From a theoretical standpoint, our formulation asymptotically enforces parsimony under the wide class of generically identifiable mixtures, which we confirmed in finite n examples.

(ii) We also showed that the required computations are no harder than for standard local priors and, although not exploited here, they are embarrassingly parallel for multiple k. We illustrated how one may simply use the output from existing MCMC algorithms for local priors, rendering the approach practical.

(iii) In particular the ECP estimator provides a convenient strategy to estimate posterior model probabilities for non-local and local priors, by utilizing readily available MCMC output and avoiding costly post-processing.

(iv) As defining prior parameters is often regarded as another practical inconvenience, here we showed how it can be advantageously calibrated to detect well-separated components resulting in multimodality.

(v) Our results showed that BIC may pathologically miss components, in some instances even with large n. The AIC and local priors tended to add spurious

components in simulations and in datasets with known subgroup structure.

(vi) In our examples the sBIC showed a mixed behavior that was similar to the BIC in some instances and to local priors or the AIC in others.

(vii) Interestingly, as an alternative to our model selection framework we attempted using overfitted and repulsive overfitted mixtures with fixed large number of components k. While these proved useful in several examples their performance can depend on tuning the Dirichlet prior parameters and potentially the choice of k. Naturally our framework can also be sensitive to prior specification, but as we illustrated there are natural default parameters based on multi-modality and minimal informativeness that result in a fairly competitive behaviour.

According to this thesis, we have the following suggestions for future work:

(i) *Parsimony enforcement.* We remark that in our examples we used a uniform prior on the model space, in future work we may achieve further parsimony by reinforcing sparse models a priori.

(ii) *Computational extensions.* An interesting venue for future research is to develop fast approximations to the two terms required to obtain the NLP integrated likelihood. For instance we could extend the approach given by (Biernacki et al., 2000) to approximate the local $p^L(\mathbf{y} \mid \mathcal{M}_k)$, and use deterministic expansions around the posterior mode $\hat{\boldsymbol{\vartheta}}_k$ to approximate the posterior expected penalty $E(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}, \mathcal{M}_k))$.

(iii) *Extensions to robust and infinite/nonparametric mixture models.* Another intriguing observation was that, by penalizing poorly-separated and low-weight components, NLPs showed robustness to model misspecification in an example. It would be interesting to study the combined effect of NLPs and robust likelihoods. Other related interesting venues are non-parametric mixtures (Murphy et al., 2017) and determinantal point processes (Xie and Xu, 2017; Bianchini et al., 2017), here we emphasize that NLPs require not only a repulsive force but also penalizing low-weight components which was found to improve inference in our examples. Particularly, as an extension of this

work, the MOM prior may be considered for the parameters of the centering distribution in the Dirichlet Process Mixture of Normal densities introducing repulsion among components through their centers, and the computations for the expected number of components may be obtained using for example a Polya urn scheme (MacEachern, 1994).

(iv) *Extensions to high dimensional settings.* In this work we consider $n \gg p$ however according to Rossell and Telesca (2017) NLPs perform well in high dimensional estimation for regular models. Therefore, an interesting avenue is the use of the MOM-IW and MOM-Beta for clustering purposes when the number of variables increase such as $n \ll p$.

Appendix A

Proofs

A.1 Auxiliary lemmas to prove Theorem 1

We state Lemma A.1.1 and Lemma A.1.2 and proof two auxiliary lemmas that will be used in the proof of Theorem 1.

Lemma A.1.1 *Let* $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = d_\theta(\boldsymbol{\theta})p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)p(\boldsymbol{\eta} \mid \mathcal{M}_k)$ *be the MOM prior in* (2.1.4). *Then* $p(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \tilde{d}_\theta(\boldsymbol{\theta})\tilde{p}^L(\boldsymbol{\theta} \mid \mathcal{M}_k)p(\boldsymbol{\eta} \mid \mathcal{M}_k)$, *where* $\tilde{d}_\theta(\boldsymbol{\theta}) \leq c_k$ *for some finite* c_k,

$$\tilde{p}^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k) = \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)gA_\Sigma\right),$$

and $\epsilon \in (0,1)$ *is an arbitrary constant.*

Proof. The MOM prior has an unbounded penalty

$$d_\theta(\boldsymbol{\theta}) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \left((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/g \right)^t,$$

however we may rewrite $d_\theta(\boldsymbol{\theta})p^L(\boldsymbol{\theta} \mid \mathcal{M}_k)$

$$\begin{aligned} &= d_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j; \mathbf{0}, gA_\Sigma\right) \frac{N\left(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)gA_\Sigma\right)}{N\left(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)gA_\Sigma\right)}, \\ &= \tilde{d}_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} N\left(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)gA_\Sigma\right), \end{aligned} \tag{A.1.1}$$

where $\epsilon \in (0,1)$ is an arbitrary constant and $\tilde{d}_\theta(\boldsymbol{\theta}) =$

$$d_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} \frac{N\left(\boldsymbol{\mu}_j; \mathbf{0}, gA_\Sigma\right)}{N\left(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)gA_\Sigma\right)} = d_\theta(\boldsymbol{\theta}) \prod_{j=1}^{k} (1+\epsilon)^{1/2} \exp\left\{-\frac{1}{2}\frac{\epsilon \boldsymbol{\mu}_j' A_\Sigma^{-1} \boldsymbol{\mu}_j}{(1+\epsilon)g}\right\}.$$

The fact that $\tilde{d}_\theta(\boldsymbol{\theta})$ is bounded follows from the fact that the product term is a Normal kernel and hence bounded, whereas $d_\theta(\boldsymbol{\theta})$ can only become unbounded when $\boldsymbol{\mu}_j A_\Sigma^{-1} \boldsymbol{\mu}_j \to \infty$ for some j, but this polynomial increase is countered by the exponential decrease in $\exp\left\{-\frac{1}{2}\frac{\epsilon \boldsymbol{\mu}_j' A_\Sigma^{-1} \boldsymbol{\mu}_j}{(1+\epsilon)g}\right\}$. □

Lemma A.1.2 *Let $d_\vartheta(\boldsymbol{\vartheta}_k) \in [0, c_k]$ be a bounded continuous function in $\boldsymbol{\vartheta}_k$, where c_k is a finite constant. Let*

$$g_k(\mathbf{y}) = E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}, \mathcal{M}_k) = \int d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k.$$

If for any $\epsilon > 0$ we have that $P^L(d_\vartheta(\boldsymbol{\vartheta}) > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 0$ then $g_k(\mathbf{y}) \xrightarrow{P} 0$. Alternatively, if there exists some $d_k^* > 0$ such that for any $\epsilon > 0$ $P^L(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 1$, then $g_k(\mathbf{y}) \xrightarrow{P} d_k^*$.

Proof. Consider the case $P^L(d_\vartheta(\boldsymbol{\vartheta}) > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 0$, then $g_k(\mathbf{y}) =$

$$\begin{aligned}
&\int_{d_\vartheta(\boldsymbol{\vartheta}_k) < \epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k + \int_{d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k \\
&\leq \epsilon P^L(d_\vartheta(\boldsymbol{\vartheta}_k) < \epsilon \mid \mathbf{y}, \mathcal{M}_k) + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \\
&\leq \epsilon + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} \epsilon,
\end{aligned}$$

where $\epsilon > 0$ is arbitrarily small. Hence $g_k(\mathbf{y}) \xrightarrow{P} 0$.

Next consider the case $P^L(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} 1$. Then

$$\begin{aligned}
g_k(\mathbf{y}) &> \int_{d_\vartheta(\boldsymbol{\vartheta}_k) > d_k^* - \epsilon} d_\vartheta(\boldsymbol{\vartheta}_k) p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}) d\boldsymbol{\vartheta}_k \\
&\geq (d_k^* - \epsilon) P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > d_k^* - \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} d_k^* - \epsilon,
\end{aligned}$$

and analogously $g_k(\mathbf{y}) =$

$$\int_{d_\vartheta(\boldsymbol{\vartheta}_k)<d_k^*+\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)d\boldsymbol{\vartheta}_k + \int_{d_\vartheta(\boldsymbol{\vartheta}_k)>d_k^*+\epsilon} d_\vartheta(\boldsymbol{\vartheta}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k)d\boldsymbol{\vartheta}_k$$
$$\leq (d_k^* + \epsilon) + c_k P^L(d_\vartheta(\boldsymbol{\vartheta}_k) > d_k^* + \epsilon \mid \mathbf{y}, \mathcal{M}_k) \xrightarrow{P} d_k^* + \epsilon,$$

for any $\epsilon > 0$ and hence $g_k(\mathbf{y}) \xrightarrow{P} d_k^*$. □

A.2 Proof of Theorem 1

Part (i). The result is straightforward. Briefly, $p(\mathbf{y} \mid \mathcal{M}_k) =$

$$\begin{aligned}
&\int d_\vartheta(\boldsymbol{\vartheta}_k)p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)d\boldsymbol{\vartheta}_k \\
&= \int d_\vartheta(\boldsymbol{\vartheta}_k)\frac{p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)}{p^L(\mathbf{y} \mid \mathcal{M}_k)}p^L(\mathbf{y} \mid \mathcal{M}_k)d\boldsymbol{\vartheta}_k \\
&= p^L(\mathbf{y} \mid \mathcal{M}_k)E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y}),
\end{aligned}$$

as desired.

Part (ii). Posterior concentration. We need to prove that

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon \mid \mathbf{y}, \mathcal{M}_k\right) \to 0$$

where $d_k^* = 0$ for $k > k^*$ and $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*)$ for $k \leq k^*$. Intuitively, the result follows from the fact that by the L_1 posterior concentration assumption B1 the posterior concentrates on the KL-optimal model $p_k^*(\mathbf{y})$, but for generically identifiable mixtures this corresponds to parameter values satisfying $d(\boldsymbol{\vartheta}_k) = 0$ if $k > k^*$ and $d(\boldsymbol{\vartheta}_k) > 0$ if $k \leq k^*$.

More formally, let A_k be the set of $\boldsymbol{\vartheta}_k \in \Theta_k$ defining $p_k^*(\mathbf{y})$, *i.e.* minimizing KL divergence between the data-generating $p(\mathbf{y} \mid \boldsymbol{\vartheta}_{k^*}^*, \mathcal{M}_{k^*})$ and $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$. Consider first the overfitted model case $k > k^*$, then generic identifiability gives that

$$A_k = \{\boldsymbol{\vartheta}_k \in \Theta_k : \eta_j = 0 \text{ for some } j = 1, \ldots, k \text{ or } \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \text{ for some } i \neq j\}.$$

This implies that for all $\boldsymbol{\vartheta}_k \in A_k$ we have that $d_\vartheta(\boldsymbol{\vartheta}_k) = 0$ and also that the L_1 distance

$$l(\boldsymbol{\vartheta}_k) = \int |p_k^*(\mathbf{y}) - p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)|\, d\mathbf{y} = 0.$$

Thus $d_\vartheta(\boldsymbol{\vartheta}_k) > 0 \Rightarrow \boldsymbol{\vartheta}_k \notin A_k \Rightarrow l(\boldsymbol{\vartheta}_k) > 0$. Given that by assumption $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k)$ are continuous in $\boldsymbol{\vartheta}_k$, for all $\epsilon' > 0$ there is an $\epsilon > 0$ such that $d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon'$ implies $l(\boldsymbol{\vartheta}_k) > \epsilon$ and hence that the probability of the former event must be smaller. That is,

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}_k) > \epsilon' \mid \mathbf{y}, \mathcal{M}_k\right) \leq P^L(l(\boldsymbol{\vartheta}_k) > \epsilon \mid \mathbf{y}, \mathcal{M}_k)$$

and the right hand side converges to 0 in probability for an arbitrary ϵ by Condition B1, proving the result for the case $k > k^*$.

The proof for the $k \leq k^*$ case proceeds analogously. Briefly, when $k \leq k^*$ generic identifiability gives that $A_k = \{\boldsymbol{\vartheta}_k^*\}$ is a singleton with positive weights $\eta_j^* > 0$ for all $j = 1, \ldots, k$ and $\boldsymbol{\theta}_i^* \neq \boldsymbol{\theta}_j^*$ for $i \neq j$. Thus $d_k^* = d_\vartheta(\boldsymbol{\vartheta}_k^*) > 0$. By continuity of $p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)$ and $d_\vartheta(\boldsymbol{\vartheta}_k)$ with respect to $\boldsymbol{\vartheta}_k$ this implies that for all $\epsilon' > 0$ there exists an $\epsilon > 0$ such that $l(\boldsymbol{\vartheta}_k) < \epsilon \Rightarrow |d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon'$, and thus that

$$P^L\left(|d_\vartheta(\boldsymbol{\vartheta}_k) - d_k^*| > \epsilon' \mid \mathbf{y}, \mathcal{M}_k\right) \leq P^L\left(l(\boldsymbol{\vartheta}_k) < \epsilon \mid \mathbf{y}, \mathcal{M}_k\right),$$

where the right hand side converges to 1 in probability by Condition B1, proving the result.

Part (ii). Convergence of $E^L(d_\vartheta(\boldsymbol{\vartheta}_k) \mid \mathbf{y})$

Consider first the case where $d_\vartheta(\boldsymbol{\vartheta}_k) \in [0, c_k]$ is bounded below some finite constant c_k. Then Part (ii) above and Lemma A.1.2 below give that

$$\begin{aligned} E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &\xrightarrow{P} 0, \text{ for } k > k^* \\ E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &\xrightarrow{P} d_k^* > 0, \text{ for } k \leq k^* \end{aligned} \tag{A.2.1}$$

as we wished to prove. Next, consider the MOM prior case (as an illustration) $d_\vartheta(\boldsymbol{\vartheta}) =$

$$d_\eta(\boldsymbol{\eta}) \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \left((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' A_\Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right),$$

where $d_\eta(\boldsymbol{\eta})$ is bounded by assumption. From Lemma A.1.1

$$\begin{aligned} E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) &= \int \tilde{d}_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta}) \frac{p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k) \tilde{p}(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)}{p^L(\mathbf{y} \mid \mathcal{M}_k)} \frac{\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k)}{\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k)} d\boldsymbol{\vartheta}_k \\ &= \frac{\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k)}{p^L(\mathbf{y} \mid \mathcal{M}_k)} \int \tilde{d}_\theta(\boldsymbol{\theta}) d_\eta(\boldsymbol{\eta}) \tilde{p}^L(\boldsymbol{\vartheta}_k \mid \mathbf{y}, \mathcal{M}_k) d\boldsymbol{\vartheta}_k, \end{aligned} \tag{A.2.2}$$

where $\tilde{d}_\theta(\boldsymbol{\theta})d_\eta(\boldsymbol{\eta})$ is bounded and hence by Part (ii) and Lemma A.1.2 the integral in (A.2.2) converges to 0 in probability when $k > k^*$ and to a non-zero finite constant when $k \leq k^*$. Therefore it suffices to show that $\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k)/p^L(\mathbf{y} \mid \mathcal{M}_k)$ is bounded in probability, as this would then immediately imply the desired result (A.2.1). From Lemma A.1.1 $\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k) =$

$$
\begin{aligned}
&\int p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)\tilde{p}^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)d\boldsymbol{\vartheta}_k = \\
&\int p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)\frac{\tilde{p}^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)}{p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)}d\boldsymbol{\vartheta}_k = \\
&\int p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)\prod_{j=1}^{k}\frac{N(\boldsymbol{\mu}_j; \mathbf{0}, (1+\epsilon)g\Sigma_j)}{N(\boldsymbol{\mu}_j; \mathbf{0}, g\Sigma_j)}d\boldsymbol{\vartheta}_k \\
&\int p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathcal{M}_k)p^L(\boldsymbol{\vartheta}_k \mid \mathcal{M}_k)\frac{1}{(1+\epsilon)^{kp/2}}\exp\left\{\frac{1}{2g}\sum_{j=1}^{k}\boldsymbol{\mu}_j' A_\Sigma^{-1}\boldsymbol{\mu}_j\frac{\epsilon}{1+\epsilon}\right\}d\boldsymbol{\vartheta}_k \\
&= \frac{p^L(\mathbf{y} \mid \mathcal{M}_k)}{(1+\epsilon)^{kp/2}}E^L\left(\exp\left\{\frac{1}{2g}\sum_{j=1}^{k}\boldsymbol{\mu}_j' A_\Sigma^{-1}\boldsymbol{\mu}_j\frac{\epsilon}{1+\epsilon}\right\} \mid \mathbf{y}, \mathcal{M}_k\right) \\
&\geq \frac{p^L(\mathbf{y} \mid \mathcal{M}_k)}{(1+\epsilon)^{kp/2}}, \qquad \text{(A.2.3)}
\end{aligned}
$$

thus $\tilde{p}^L(\mathbf{y} \mid \mathcal{M}_k)/p^L(\mathbf{y} \mid \mathcal{M}_k) \geq \frac{1}{(1+\epsilon)^{kp/2}}$. From (A.2.2) this implies that when $k \leq k^*$ we obtain $E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} d_k^* > 0$. Further, by Condition B3 the $E^L()$ term in (A.2.3) is bounded above in probability when $k > k^*$, implying that $E^L\left(d_\vartheta(\boldsymbol{\vartheta}) \mid \mathbf{y}, \mathcal{M}_k\right) \xrightarrow{P} 0$. □

Part (iii).

By assumption $p(\boldsymbol{\eta} \mid \mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta}; q) \propto d_\eta(\boldsymbol{\eta})\mathrm{Dir}(\boldsymbol{\eta}; q-r)$, where $d_\eta(\boldsymbol{\eta}) = \prod_{j=1}^{k}\eta_j^r$ and $q > 1$, $q - r < 1$. Consider the particular choice $q - r < \dim(\Theta)/2$ and without loss of generality let $k^*+1, \ldots, k$ be the labels for the spurious components. Theorem 1 in Rousseau and Mengersen (2011) showed that under the assumed A1-A4 and a further condition A5 trivially satisfied by $p^L(\boldsymbol{\eta} \mid \mathcal{M}_k) = \mathrm{Dir}(\boldsymbol{\eta}; q-r)$ the corresponding posterior distribution of the spurious weights concentrates around 0, specifically

$$
P^L\left(\sum_{j=k^*+1}^{k}\eta_j > n^{-\frac{1}{2}+\tilde{\epsilon}} \mid \mathbf{y}, \mathcal{M}_k\right) \to 0 \qquad \text{(A.2.4)}
$$

in probability for all $\tilde{\epsilon} > 0$ as $n \to \infty$. Now, the fact that the geometric mean is

smaller than the arithmetic mean gives that

$$(k-k^*)\left(\prod_{j=k^*+1}^{k}\eta_j\right)^{\frac{1}{k-k^*}} \leq \sum_{j=k^*+1}^{k}\eta_j,$$

and thus

$$P^L\left(\sum_{j=k^*+1}^{k}\eta_j > n^{-\frac{1}{2}+\tilde{\epsilon}} \mid \mathbf{y}, \mathcal{M}_k\right) \geq$$

$$P^L\left((k-k^*)\left(\prod_{j=k^*+1}^{k}\eta_j\right)^{\frac{1}{k-k^*}} > n^{-\frac{1}{2}+\tilde{\epsilon}} \mid \mathbf{y}, \mathcal{M}_k\right) =$$

$$P^L\left(\prod_{j=k^*+1}^{k}\eta_j^r > \frac{1}{(k-k^*)^r}n^{-\frac{r(k-k^*)}{2}+\epsilon} \mid \mathbf{y}, \mathcal{M}_k\right), \tag{A.2.5}$$

where $\epsilon = r(k-k^*)\tilde{\epsilon}$ is a constant. Thus (A.2.4) implies that (A.2.5) also converges to 0 in probability. Finally, given that by assumption $d_\vartheta(\boldsymbol{\vartheta}) = d_\theta(\boldsymbol{\theta})d_\eta(\boldsymbol{\eta}) \leq c_k\prod_{j=k^*+1}^{k}\eta_j^r$ we obtain

$$P^L\left(d_\vartheta(\boldsymbol{\vartheta}) > n^{-\frac{r(k-k^*)}{2}+\epsilon} \mid \mathbf{y}, \mathcal{M}_k\right) \leq P^L\left(\prod_{j=k^*+1}^{k}\eta_j^r > \frac{1}{c_k}n^{-\frac{r(k-k^*)}{2}+\epsilon} \mid \mathbf{y}, \mathcal{M}_k\right), \tag{A.2.6}$$

where the right hand side converges in probability to 0 given that (A.2.5) converges to 0 in probability and c_k, k, k^*, r are finite constants. As mentioned earlier this result holds for any $r > 0$ satisfying $q - r < \dim(\Theta)/2$, in particular we may set $q - r = \delta < \dim(\Theta)/2$ (where $\delta > 0$ can be arbitrarily small) so that plugging $r = q - \delta$ into the left hand side of (A.2.6) gives the desired result. □

A.3 Proof of Lemma 1

Let D_{ij} be a $pk \times pk$ matrix where the i^{th} and j^{th} diagonal blocks are equal to the $p \times p$ identity matrix, and the (i,j) off-diagonal block is minus the identity matrix, so that $(\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j)'(\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j) = \boldsymbol{\zeta}' D_{ij}\boldsymbol{\zeta}$. Then a direct application of Lemma 1 in Kan

(2006) gives that

$$d_k(\boldsymbol{\zeta}) = \prod_{i<j}(\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j)'(\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j) = \prod_{i<j}\boldsymbol{\theta}' D_{ij}\boldsymbol{\zeta} =$$

$$= \frac{1}{[k(k-1)/2]!}\sum_{v(1,2)=0}^{1}\sum_{v(k-1,k)=0}^{1}(-1)^{\sum_{i<j}v(i,j)}\left[\boldsymbol{\zeta}'\left(\sum_{i<j}\left(\frac{1}{2}-v(i,j)\right)D_{ij}\right)\boldsymbol{\zeta}\right]^{\frac{k(k-1)}{2}}$$

$$= \frac{1}{[k(k-1)/2]!}\sum_{v(1,2)=0}^{1}\sum_{v(k-1,k)=0}^{1}(-1)^{\sum_{i<j}v(i,j)}\left[\boldsymbol{\zeta}' B_v \boldsymbol{\zeta}\right]^{\frac{k(k-1)}{2}} \tag{A.3.1}$$

where $B_v = \left(\sum_{i<j}\left(\frac{1}{2} - v(i,j)\right)D_{ij}\right)$ is a matrix with element (l,m) given by

$$\begin{cases} b_{ll} = \frac{1}{2}(k-1) - \sum_{i<j}v(i,j),\ l = 1+p(i-1),\ldots,pi \\ b_{lm} = b_{ml} = -\frac{1}{2} + \sum_{i<j}v(i,j),\ (1+p(i-1), 1+p(j-1)),\ldots,(pi,pj) \end{cases}.$$

Let ζ_l be the l^{th} element in $\boldsymbol{\zeta}$, then following Expression (6.1) in (Mohsenipour, 2012)

$$\left[\boldsymbol{\zeta}' B_v \boldsymbol{\zeta}\right]^{\frac{k(k-1)}{2}} = \sum_{s\in S_k}[k(k-1)/2]!\left(\prod_{l=1}^{pk}\prod_{m=1}^{pk}\frac{b_{lm}^{s_{lm}}}{s_{lm}!}\right)\prod_{l=1}^{pk}\zeta_l^{\sum_{m=1}^{pk}s_{lm}+s_{ml}} \tag{A.3.2}$$

where $s = (s_{1,1}, s_{1,2}, \ldots, s_{pk,pk})$ is a $(pk)^2$ integer vector, S_k denotes the set of partitions of $k(k-1)/2$ such that $\sum_{l=1}^{pk}\sum_{m=1}^{pk}s_{l,m} = k(k-1)/2$ with $0 \leq s_{l,m} \leq k(k-1)/2$. Plugging (A.3.2) into (A.3.1) gives that the prior normalization constant is

$$E^L(d_k(\boldsymbol{\zeta})) = \sum_{v(1,2)=0}^{1}\sum_{v(k-1,k)=0}^{1}(-1)^{\sum_{i<j}v(i,j)}\sum_{s\in S_k}\left(\prod_{l=1}^{pk}\prod_{m=1}^{pk}\frac{b_{lm}^{s_{lm}}}{s_{lm}!}\right)\prod_{l=1}^{pk}\kappa_{sl} \tag{A.3.3}$$

where $\kappa_{sl} = E^L(\zeta_{jf}^{\sum_{m=1}^{pk}s_{lm}+s_{ml}})$. □

A.4 Proof of Corollary 1

In order to compute the normalization, C_k we need to find the expectation:

$$C_k = E\left(\prod_{1\leq i<j\leq k}\left(\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^{'}A_{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{g}\right)\right).$$

with respect to $(\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_k \sim N(\mathbf{0},A_{\Sigma})$. Moreover consider the Cholesky decomposition $A_{\Sigma}=\boldsymbol{L}\boldsymbol{L}^{'}$ where $A_{\Sigma}^{-1}=(\boldsymbol{L}^{'})^{-1}\boldsymbol{L}^{-1}$, by setting $\sqrt{g}\boldsymbol{L}\boldsymbol{\mu}_j^{*}=\boldsymbol{\mu}_j$ the jacobian of the transformation is the determinant of the block diagonal matrix:

$$|J(\boldsymbol{\mu}_1^{*},...,\boldsymbol{\mu}_k^{*})|=\left|\begin{pmatrix}\sqrt{g}\boldsymbol{L} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{g}\boldsymbol{L}\end{pmatrix}\right|=g^{k/2}(\det(\boldsymbol{L}))^{k},$$

where $(\det(\boldsymbol{L}))^{k}=(\det(A_{\Sigma}))^{k/2}$. The normalization constant C_k can be found by using the following expectation

$$C_k = E\left(\prod_{1\leq i<j\leq k}((\boldsymbol{\mu}_i^{*}-\boldsymbol{\mu}_j^{*})^{'}(\boldsymbol{\mu}_i^{*}-\boldsymbol{\mu}_j^{*}))\right), \tag{A.4.1}$$

where $\boldsymbol{\mu}_k^{*}\sim N_p\left(\boldsymbol{\mu}_k^{*};\mathbf{0},\boldsymbol{I_p}\right)$.

To obtain the result we apply the adapted Proposition 4 in Kan (2006) to the $p\times k$ vector $\boldsymbol{\mu}^{*}=(\boldsymbol{\mu}_1^{*},...,\boldsymbol{\mu}_k^{*})$, where k is the number of components and $\boldsymbol{\mu}_j^{*}\in\mathbb{R}^{p}$ for $j=1,\ldots,k$, which for convenience we reproduce below as Proposition 1.

Proposition 1 *Suppose* $\boldsymbol{\mu}^{*}=(\mu_1^{*},...,\mu_k^{*})^{'}\sim N_k(\mathbf{0},\boldsymbol{I}_k)$, *for symmetric matrices* $A_{(1,2)},...,A_{(k-1,k)}$, *we have*

$$E\left(\prod_{1\leq i<j\leq k}(\boldsymbol{\mu}^{*'}A_{(i,j)}\boldsymbol{\mu}^{*})\right)=\frac{1}{s!}\sum_{\upsilon_{(1,2)}=0}^{1}...\sum_{\upsilon_{(k-1,k)}=0}^{1}(-1)^{\sum_{i,j}^{\binom{k}{2}}\upsilon_{(i,j)}}\mathcal{Q}_s(B_{\upsilon}), \tag{A.4.2}$$

where $s=\binom{k}{2}$, $B_{\upsilon}=(\frac{1}{2}-\upsilon_{(1,2)})A_{(1,2)}+,...,+(\frac{1}{2}-\upsilon_{(k-1,k)})A_{(k-1,k)}$ *and* $\mathcal{Q}_s(B_{\upsilon})$ *is given by the recursive equation:* $\mathcal{Q}_s(B_{\upsilon})=s!2^{s}d_s(B_{\upsilon})$ *where* $d_s(B_{\upsilon})=\frac{1}{2s}\sum_{i=1}^{s}tr(B_{\upsilon}^{i})d_{s-i}(B_{\upsilon})$ *and* $d_0(B_{\upsilon})=1$ *and* $A_{(i,j)}$ *is a* $pk\times pk$ *matrix* (l,m) *element*

$$\begin{cases} a_{ll}=1, \quad l=1+p(i-1)...p_i \quad and \quad l=1+p(j-1)...p_j.\\ a_{lm}=a_{ml}=-1, \quad (l,m)=(1+p(i-1),1+p(j-1))...(pi,pj).\\ a_{lm}=0 \quad otherwise.\end{cases}$$

We define now the $A_{(1,2)}, ..., A_{(k-1,k)}$ matrices with dimensions $pk \times pk$. These matrices can be found using $p * p$ identity matrices in the diagonal blocks corresponding to the i and j components minus the identity matrix in the "cross-blocks" corresponding to (i, j). Finally using the $A_{(i,j)}$ matrices, B_υ can be expressed as a $pk \times pk$ matrix with element (l, m) as follows

$$\begin{cases} b_{ll} = \frac{1}{2}(k-1) - \sum_{i<j} \upsilon_{(i,j)}, \quad l = 1 + p(i-1)...p_i \quad \text{and} \quad l = 1 + p(j-1)...p_j. \\ b_{lm} = b_{ml} = -\frac{1}{2} + \sum_{i<j} \upsilon_{(i,j)}, \quad (l, m) = (1 + p(i-1), 1 + p(j-1))...(pi, pj). \end{cases}$$

□

A.5 Proof of Corollary 2

Using the Corollary 2.2 in Lu and Richards (1993), if $z > -1/n$, then

$$(2\pi)^{-n/2} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{1 \le i < j \le n} (x_i - x_j)^{2z} \prod_{j=1}^{n} \exp\{-x_j^2/2\} dx_j = \prod_{j=1}^{n} \frac{\Gamma(jz+1)}{\Gamma(z+1)}, \tag{A.5.1}$$

and using $x_i = (\mu_i - m)/(\sqrt{a_{\sigma^2} g})$ with $i = 1, ..., k$, we have that the normalization constant for a Normal mixture ($p = 1$) is

$$C_k = E_{\mu_1,...,\mu_k} \mid a_{\sigma^2} \left(\prod_{1 \le i < j \le k} \left(\frac{\mu_i - \mu_j}{\sqrt{a_{\sigma^2} g}} \right)^{2t} \right) = \prod_{j=1}^{k} \frac{\Gamma(jt+1)}{\Gamma(t+1)}, \tag{A.5.2}$$

and for $k = 2$ is straightforward to show that $C_k = E(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\mu_i - \boldsymbol{\mu}_j) = 2tr(I_p)$. This Corollary can be generalizable to MOM-Gamma priors with $p = 1$ using Lemma 3.3 in Lu and Richards (1993).

□

A.6 Proof of Corollary 3

For $p = 1$ C_k is computed using (3.10) in Lu and Richards (1993) and for $k = 2$ is straightforward to show that $C_k = E(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) = 2 \sum_{f=1}^{p} V(\boldsymbol{\theta}_{jf})$. □

A.7 Proof of Proposition 2

We start by noting that

$$p(\mathbf{y} \mid \mathcal{M}_k) = \sum_{\mathbf{z}:n_k=0} p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_k)p(\mathbf{z} \mid \mathcal{M}_k) + \sum_{\mathbf{z}:n_k>1} p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_k)p(\mathbf{z} \mid \mathcal{M}_k) \quad \text{(A.7.1)}$$

From C1, for any $\mathbf{z}$ such that $n_k = 0$ we have that $p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_k) =$

$$\int p(\mathbf{y} \mid \boldsymbol{\vartheta}_k, \mathbf{z}, \mathcal{M}_k)p(\boldsymbol{\vartheta}_k \mid \mathbf{z}, \mathcal{M}_k)d\boldsymbol{\vartheta}_k = \int \left(\prod_{j=1}^{k-1} \prod_{z_i=j} p(\mathbf{y}_i \mid \boldsymbol{\theta}_j)\right) p(\boldsymbol{\vartheta}_k \mid \mathbf{z}, \mathcal{M}_k)d\boldsymbol{\vartheta}_k =$$
$$\int \left(\prod_{j=1}^{k-1} \prod_{z_i=j} p(\mathbf{y}_i \mid \boldsymbol{\theta}_j)\right) p(\boldsymbol{\vartheta}_{k-1} \mid \mathbf{z}, \mathcal{M}_{k-1})d\boldsymbol{\vartheta}_{k-1} = p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_{k-1}) \quad \text{(A.7.2)}$$

where the second line in (A.7.2) follows from C4. Further, from Condition C3, for any $\mathbf{z}$ such that $n_k = 0$ we have

$$p(\mathbf{z} \mid \mathcal{M}_{k-1}) = p(\mathbf{z} \mid n_k = 0, \mathcal{M}_k) = \frac{p(\mathbf{z} \mid \mathcal{M}_k)}{P(n_k = 0 \mid \mathcal{M}_k)} \Rightarrow p(\mathbf{z} \mid \mathcal{M}_k) = p(\mathbf{z} \mid \mathcal{M}_{k-1})P(n_k = 0 \mid \mathcal{M}_k). \quad \text{(A.7.3)}$$

Plugging (A.7.2) and (A.7.3) into (A.7.1) gives that $p(\mathbf{y} \mid \mathcal{M}_k) =$

$$P(n_k = 0 \mid \mathcal{M}_k) \sum_{\mathbf{z}:n_k=0} p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_{k-1})p(\mathbf{z} \mid \mathcal{M}_{k-1}) + \sum_{\mathbf{z}:n_k>1} p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_k)p(\mathbf{z} \mid \mathcal{M}_k) =$$
$$P(n_k = 0 \mid \mathcal{M}_k)p(\mathbf{y} \mid \mathcal{M}_{k-1}) + \sum_{\mathbf{z}:n_k>1} p(\mathbf{y} \mid \mathbf{z}, \mathcal{M}_k)p(\mathbf{z} \mid \mathcal{M}_k) \quad \text{(A.7.4)}$$

That is, $p(\mathbf{y} \mid \mathcal{M}_k)$ is a linear combination of $p(\mathbf{y} \mid \mathcal{M}_{k-1})$ and a sum of $p(\mathbf{y}, \mathbf{z} \mid \mathcal{M}_k)$ over cluster configurations such that the last cluster k is occupied. This recursive relation is an extension of Theorem 3.1 in Nobile (2004), who proved a similar result under more restrictive conditions than our C1-C4. Dividing both sides of (A.7.4) by $p(\mathbf{y} \mid \mathcal{M}_k)$ and rearranging terms gives

$$B_{k-1,k}(\mathbf{y}) = \frac{1}{P(n_k = 0 \mid \mathcal{M}_k)} \left(1 - \sum_{\mathbf{z}:n_k>1} \frac{p(\mathbf{y}, \mathbf{z} \mid \mathcal{M}_k)}{p(\mathbf{y} \mid \mathcal{M}_k)}\right) = \frac{P(n_k = 0 \mid \mathbf{y}, \mathcal{M}_k)}{P(n_k = 0 \mid \mathcal{M}_k)}.$$

Finally, from Condition C2 both the likelihood and prior are invariant to label

permutations and thus $P(n_j = 0 \mid \mathbf{y}, \mathcal{M}_k) = P(n_k = 0 \mid \mathbf{y}, \mathcal{M}_k)$ for any $j \neq k$, hence

$$B_{k-1,k}(\mathbf{y}) = \frac{1}{kP(n_k = 0 \mid \mathcal{M}_k)} \sum_{j=1}^{k} P(n_j = 0 \mid \mathbf{y}, \mathcal{M}_k),$$

as we wished to prove. □

For completeness we derive $P(n_k = 0 \mid \mathcal{M}_k)$ when $\boldsymbol{\eta} \sim \mathrm{Dir}(q)$. From (A.7.3), $P(n_k = 0 \mid \mathcal{M}_k) =$

$$\frac{p(\mathbf{z} \mid \mathcal{M}_k)}{p(\mathbf{z} \mid \mathcal{M}_{k-1})} = \frac{\Gamma(kq) \prod_{j=1}^{k} \Gamma(n_j + q)}{\Gamma(q)^k \Gamma(n + kq)} \frac{\Gamma(q)^{k-1} \Gamma(n + (k-1)q)}{\Gamma((k-1)q) \prod_{j=1}^{k-1} \Gamma(n_j + q)} = \frac{\Gamma(kq)\Gamma(n + (k-1)q)}{\Gamma(n + kq)\Gamma((k-1)q)}$$

Appendix B

MCMC results

In this section we assessed MCMC convergence for the considered data sets after a burn-in period via MCMC iteration plots. In Figures B.1-B.4 we consider the Old Faithful, misspecified example, Fisher's Iris and Graf-versus-Host flow cytometry data sets of Sections 5.2 and 7.1-7.3. The top panels illustrate the means μ_{jf} where the colours in the trace plots indicate the different dimensions $f = 1, ..., p$. The middle and bottom panels in Figures B.1-B.7 show the variance and weight parameters. Figure B.8 show the weight parameters for the example considered in Section 5.5. Finally, Figures B.9-B.13 illustrate the mean and weight parameters for the political blog data (see Section 7.5) where the colours in the trace plots indicate the two components. The plots do not reveal issues with the mixing and the obtained estimates were fairly close to the simulation truth in the misspecified example and the illustration of computations under product of Binomial mixtures in Section 5.5.

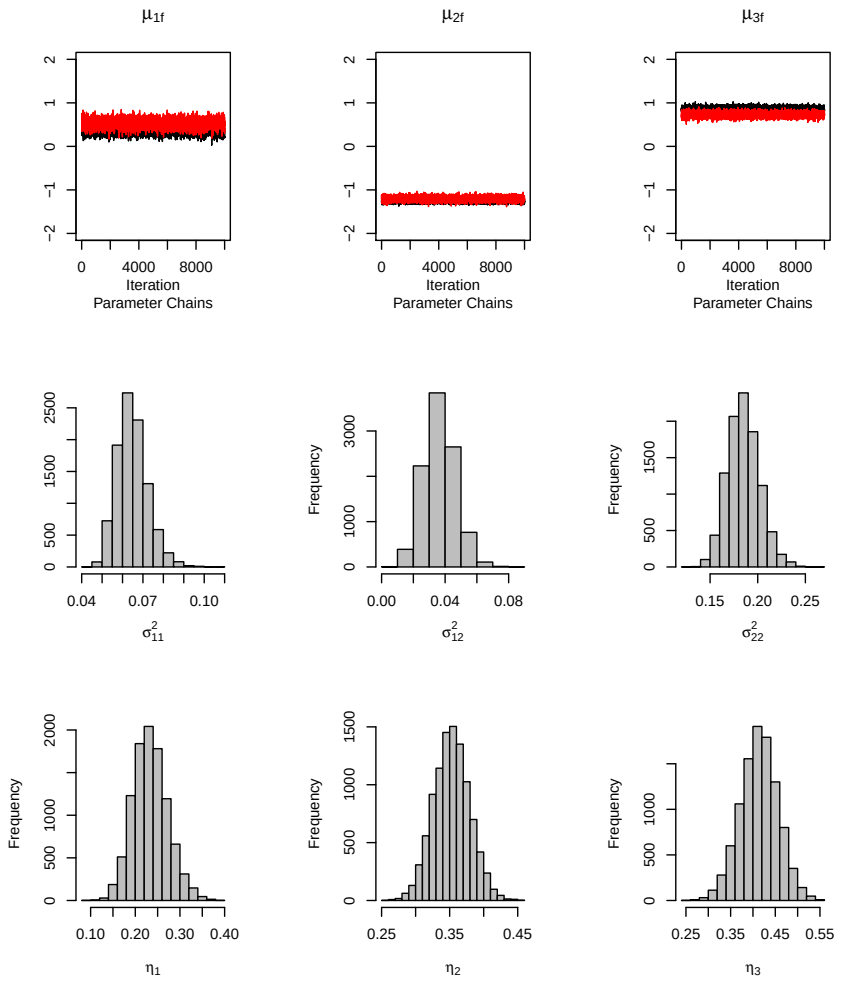


Figure B.1: MCMC results for the faithful data set with 20000 iterations and a 10000 burning period. Top: Trace plots for the mean parameters. Middle: MCMC output for variance parameters. Bottom: MCMC output for weight parameters.

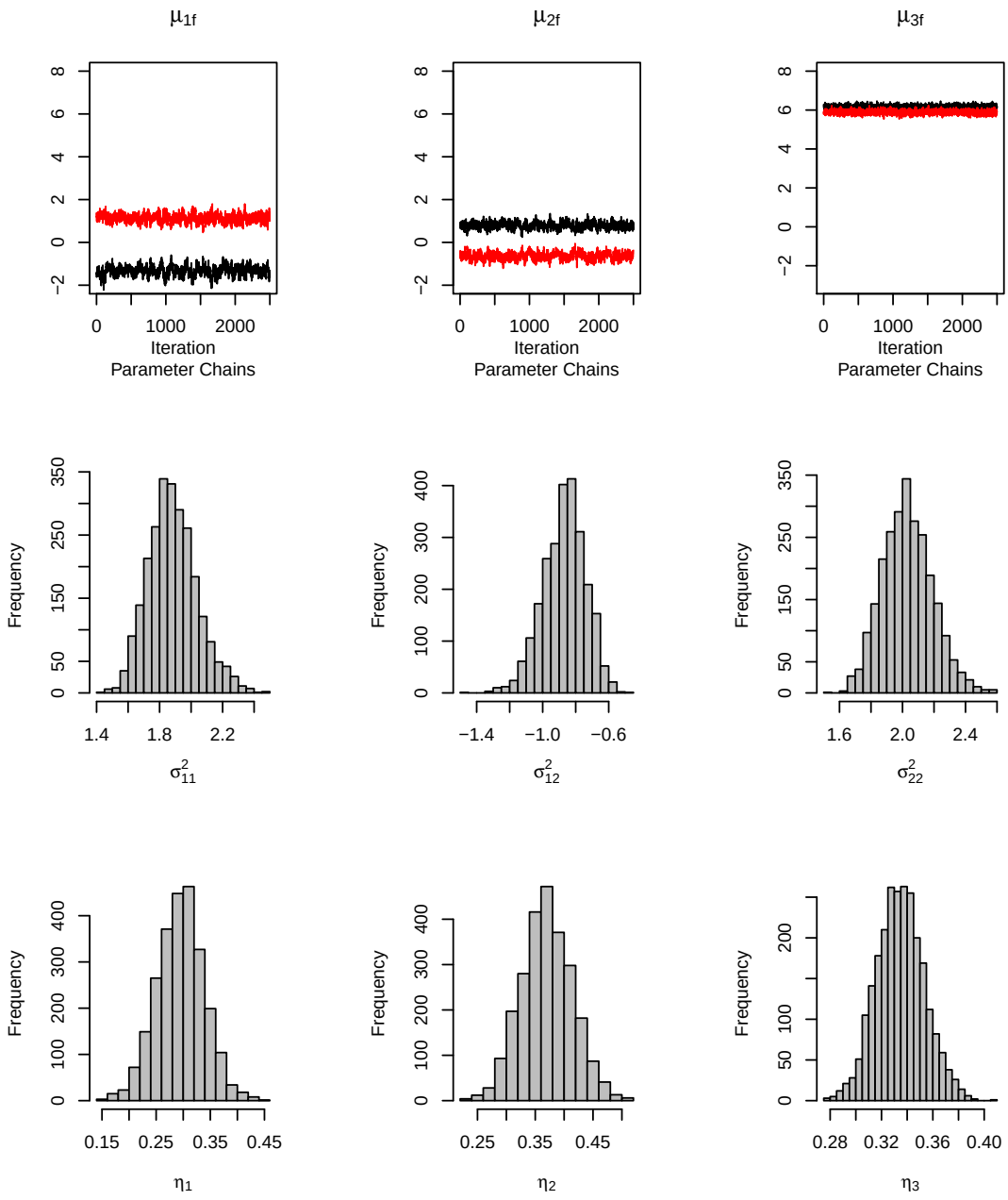


Figure B.2: MCMC results for the misspecified data set with 5000 iterations and a 2500 burning period. Top: Trace plots for the mean parameters. Middle: MCMC output for variance parameters. Bottom: MCMC output for weight parameters.

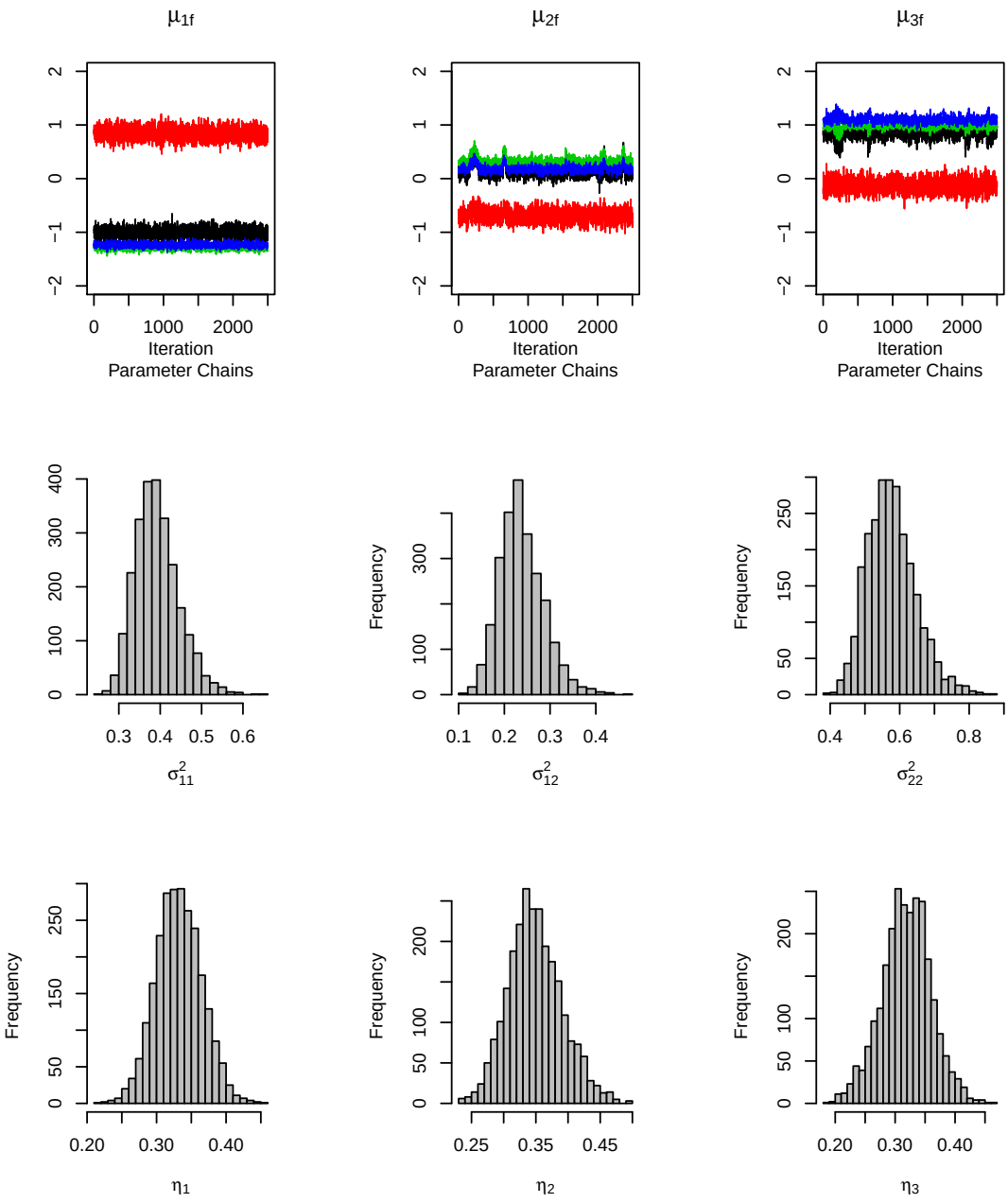


Figure B.3: MCMC results for the Iris data set with 5000 iterations and a 2500 burning period. Top: Trace plots for the mean parameters. Middle: MCMC output for variance parameters. Bottom: MCMC output for weight parameters.

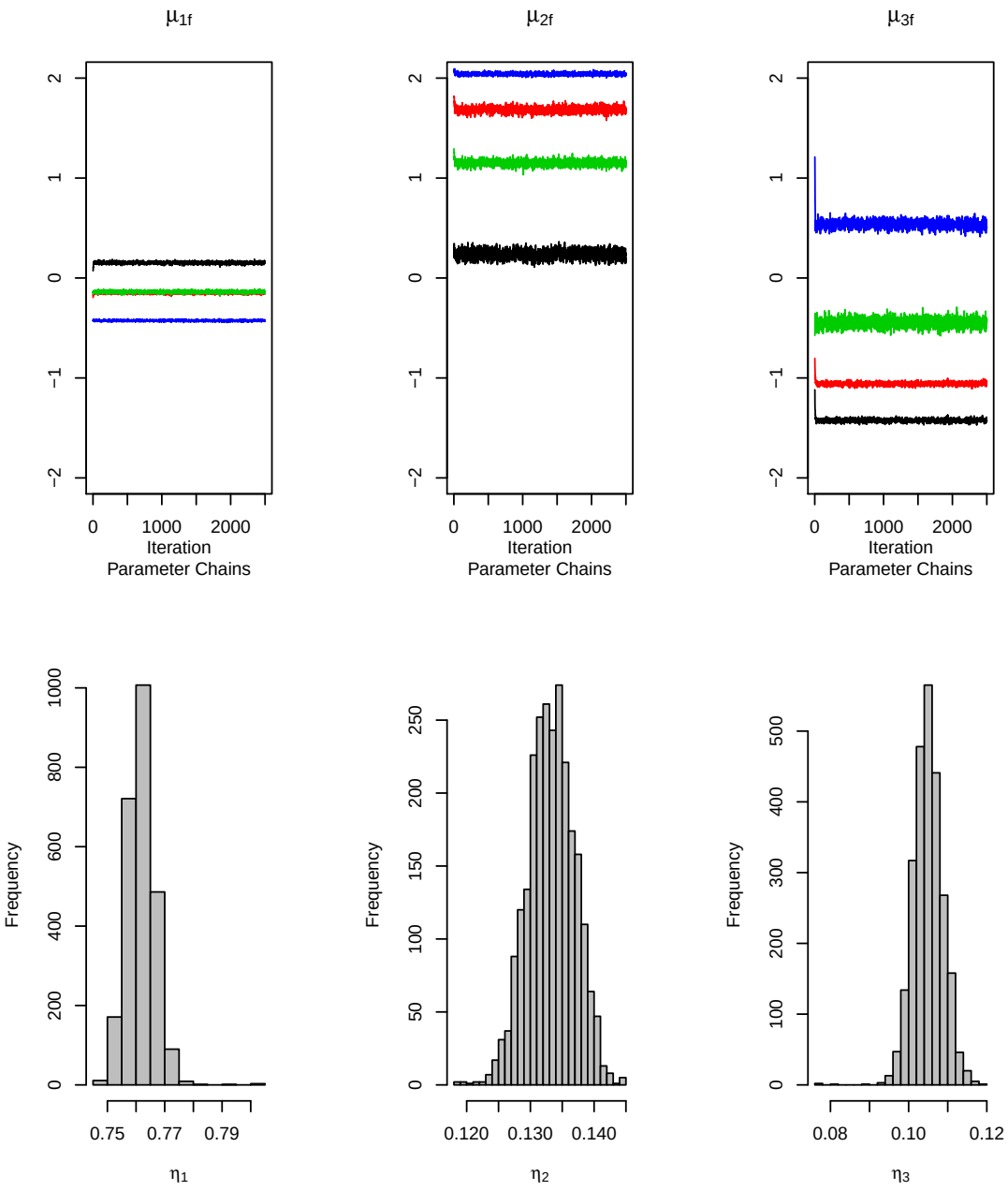


Figure B.4: MCMC results for the Cytometry data set with 5000 iterations and a 2500 burning period. Top: Trace plots for the mean parameters. Bottom: MCMC output for weight parameters.

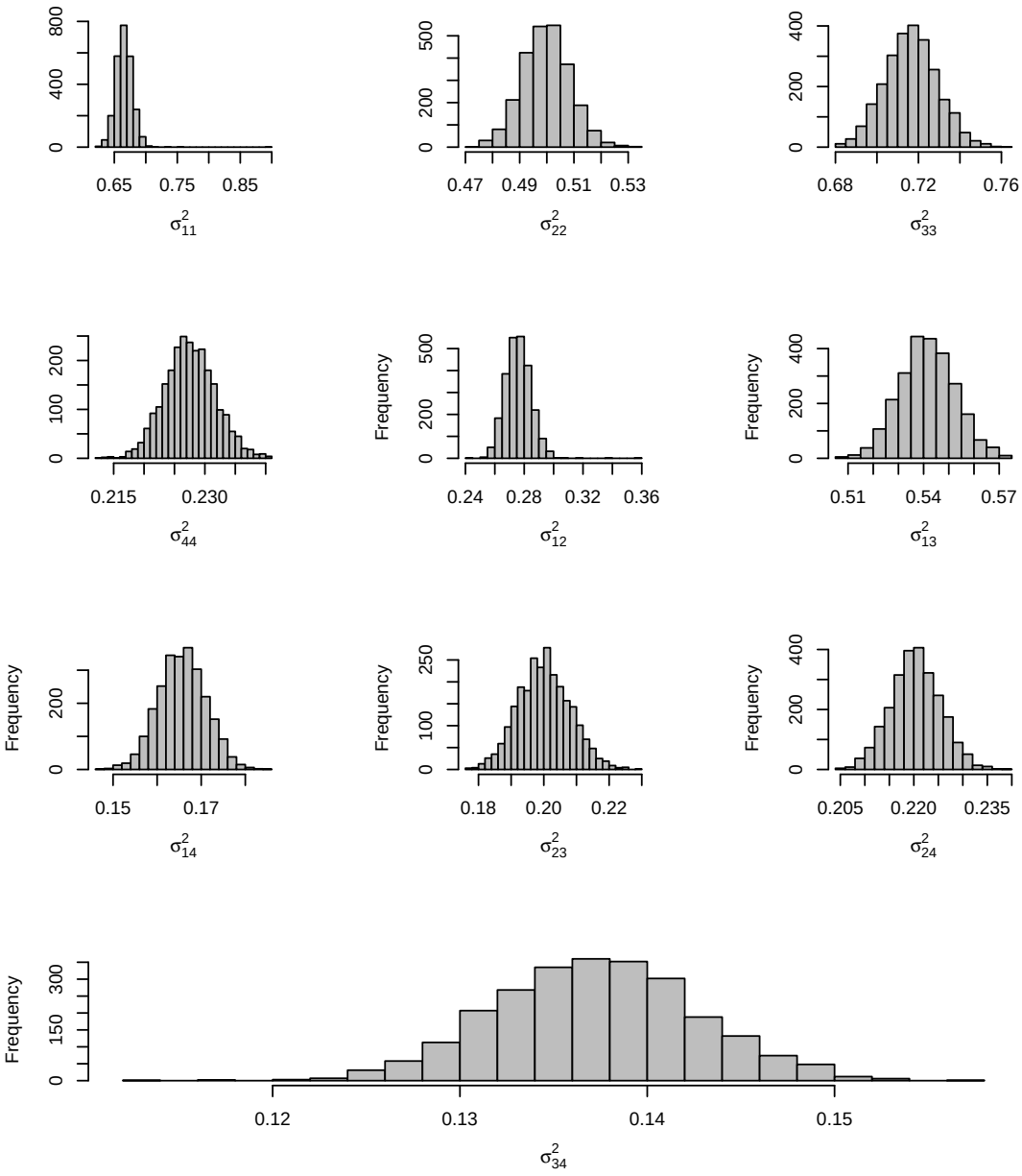


Figure B.5: MCMC results for the Cytometry data set with 5000 iterations and a 2500 burning period. MCMC output for variance parameters of the first component.

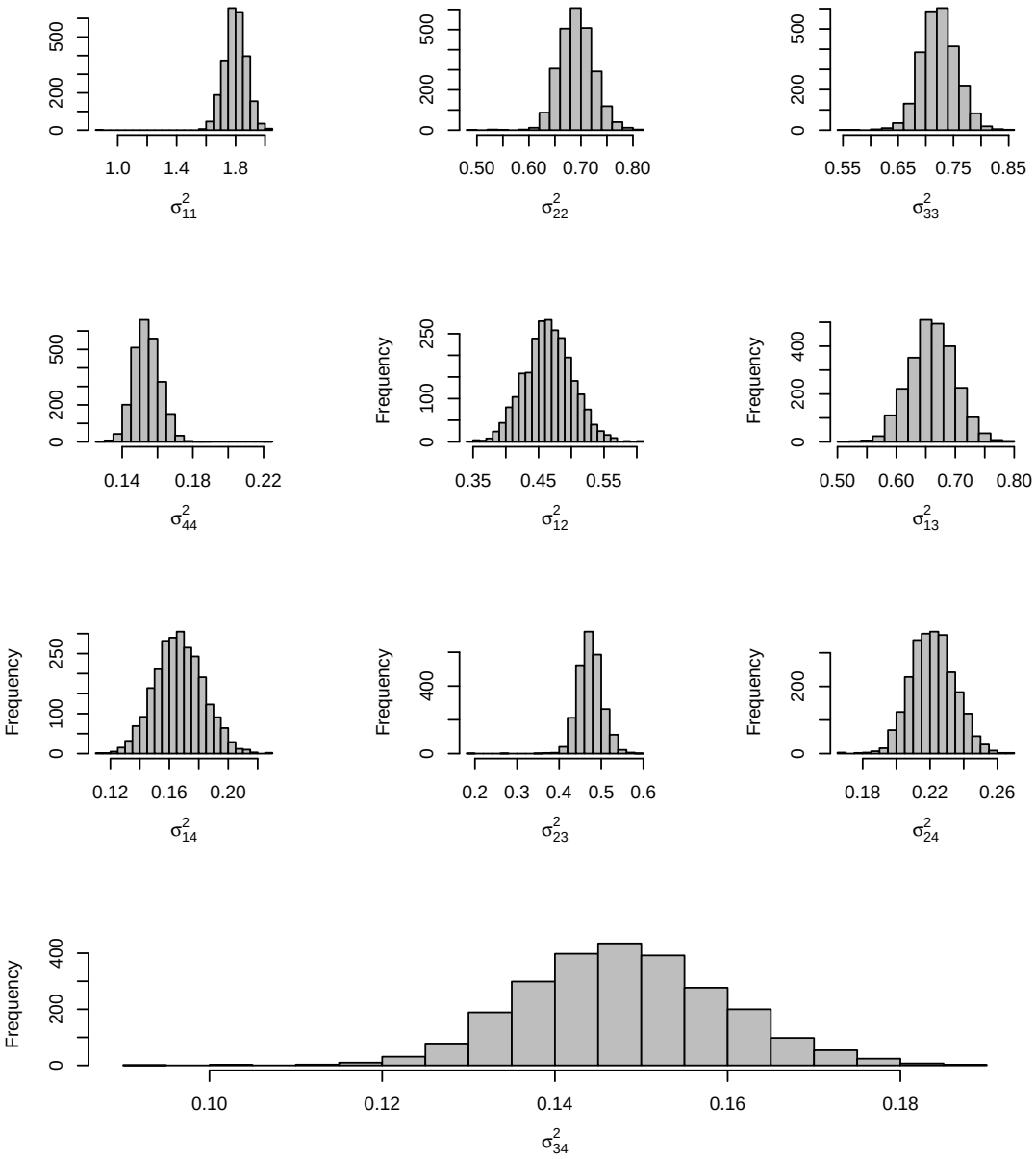


Figure B.6: MCMC results for the Cytometry data set with 5000 iterations and a 2500 burning period. MCMC output for variance parameters of the second component.

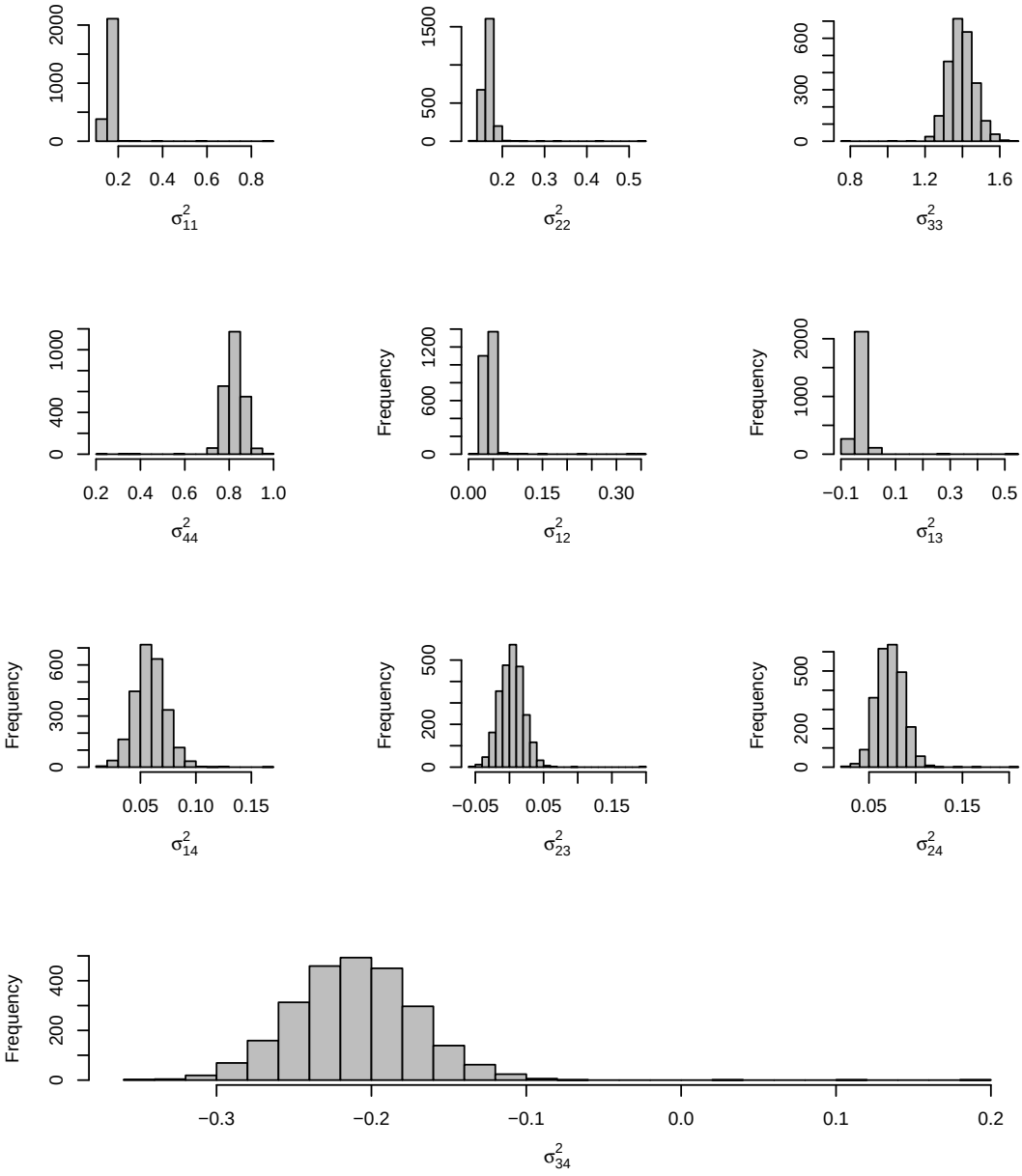


Figure B.7: MCMC results for the Cytometry data set with 5000 iterations and a 2500 burning period. MCMC output for variance parameters of the third component.

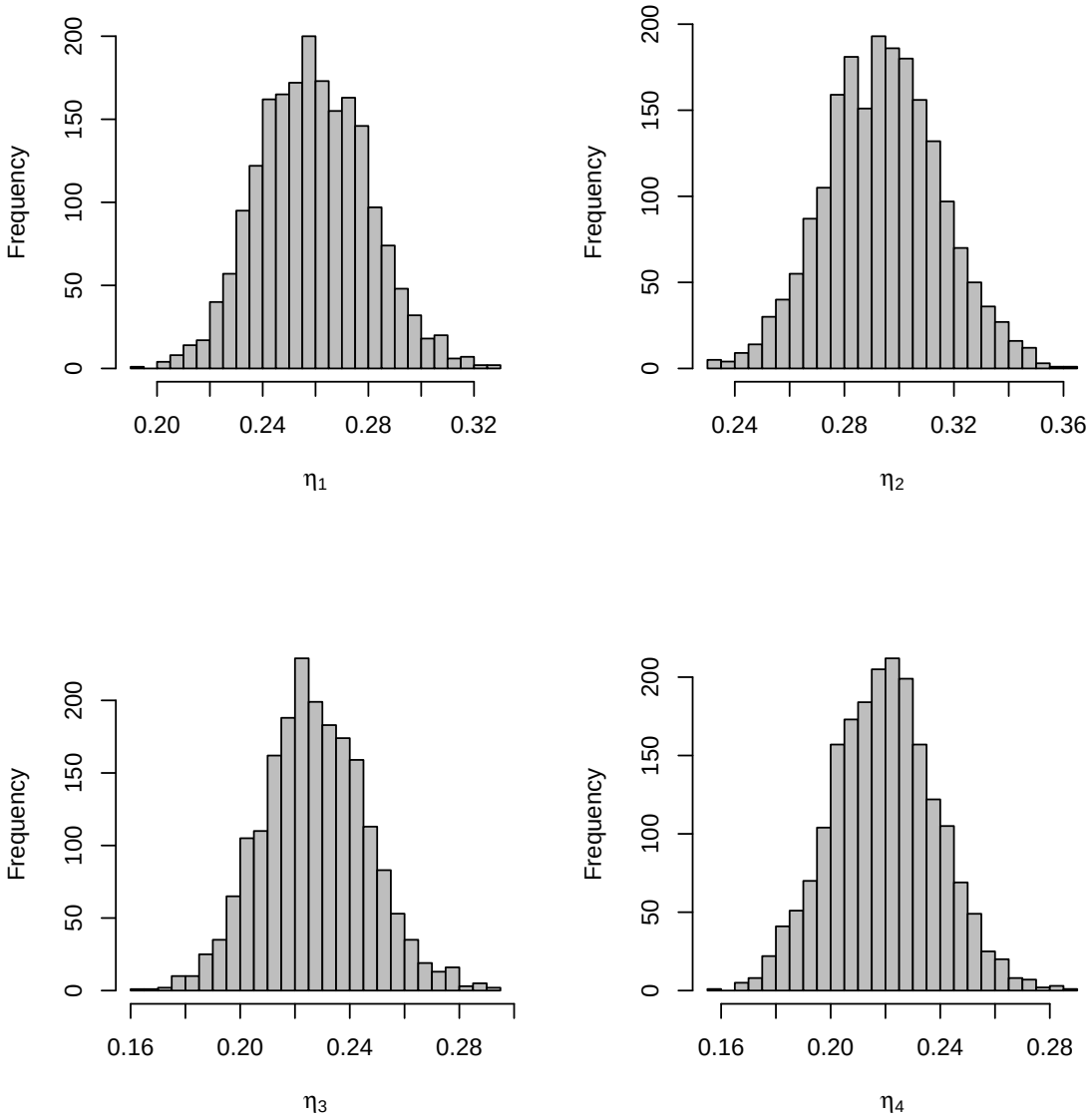


Figure B.8: Additional MCMC results for the computations for the product of Binomial mixture under MOM-Beta priors. MCMC output for weight parameters.

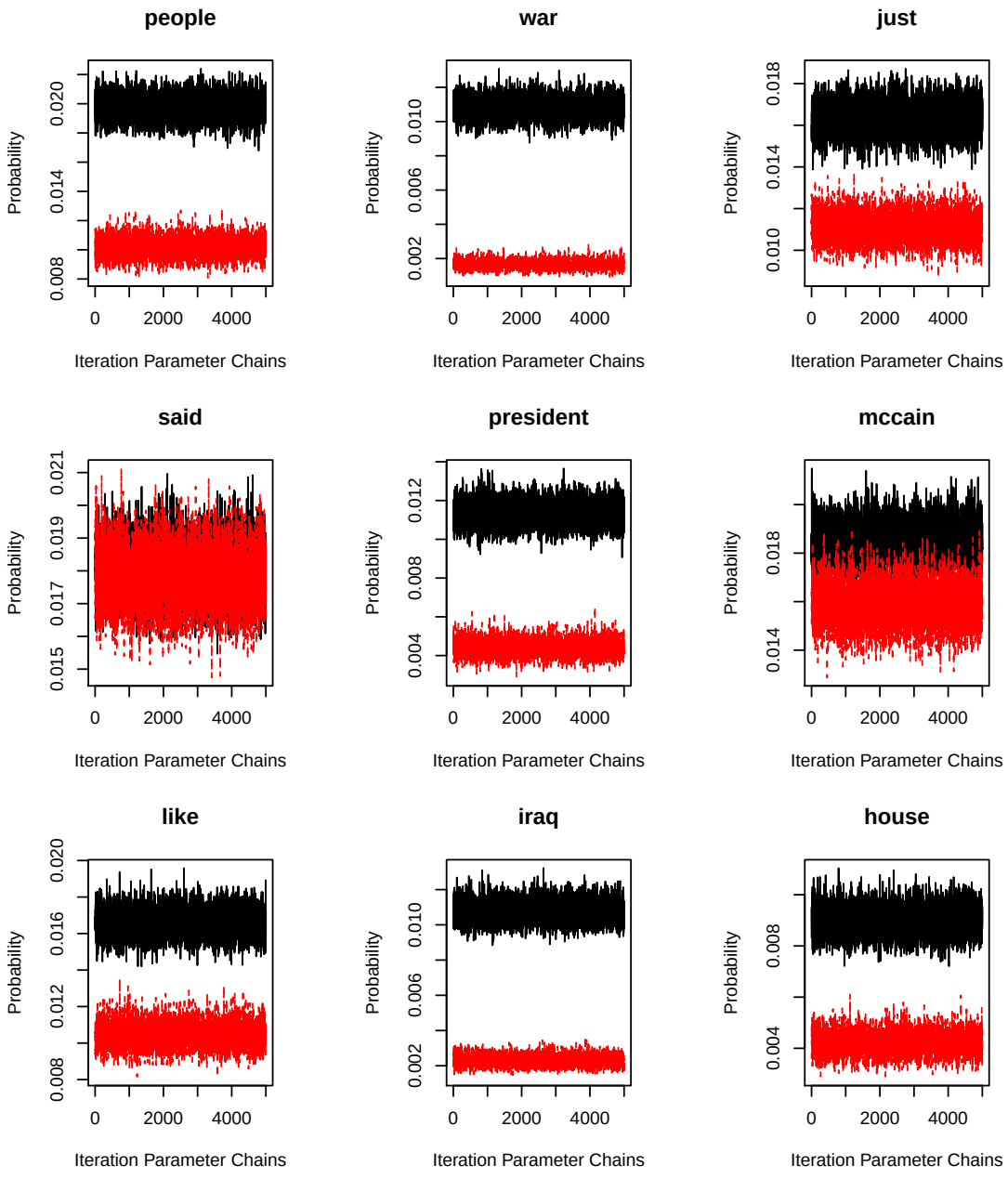


Figure B.9: MCMC results for the Political blog data with 10000 iterations, a 5000 burning period and a thinning of 10 iterations. MCMC output for the words, $\hat{\theta}_{jf}$, presented in Table 7.7.

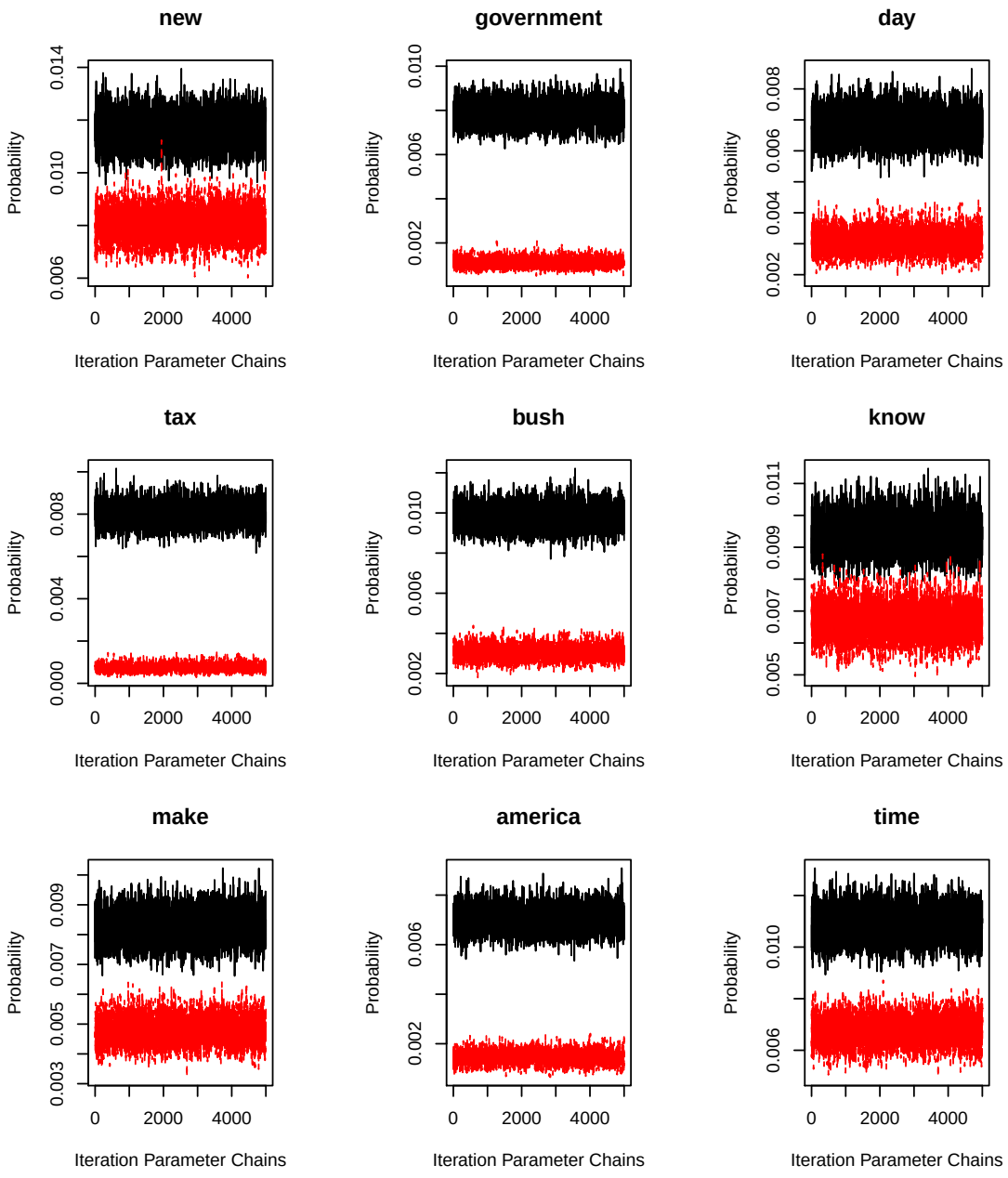


Figure B.10: MCMC results for the Political blog data with 10000 iterations, a 5000 burning period and a thinning of 10 iterations. MCMC output for the words, $\hat{\theta}_{jf}$, presented in Table 7.7.

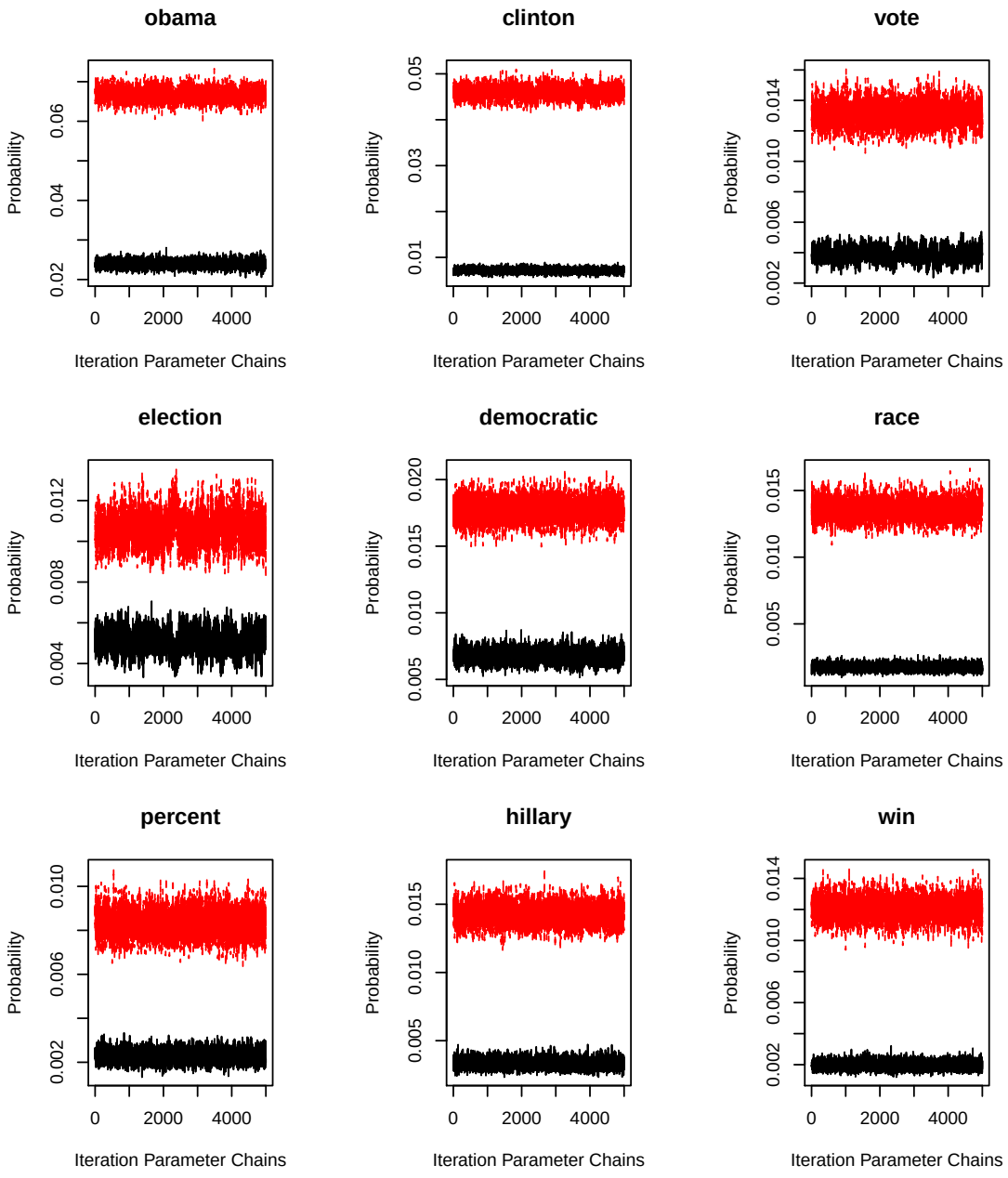


Figure B.11: MCMC results for the Political blog data with 10000 iterations, a 5000 burning period and a thinning of 10 iterations. MCMC output for the words, $\hat{\theta}_{jf}$, presented in Table 7.7.

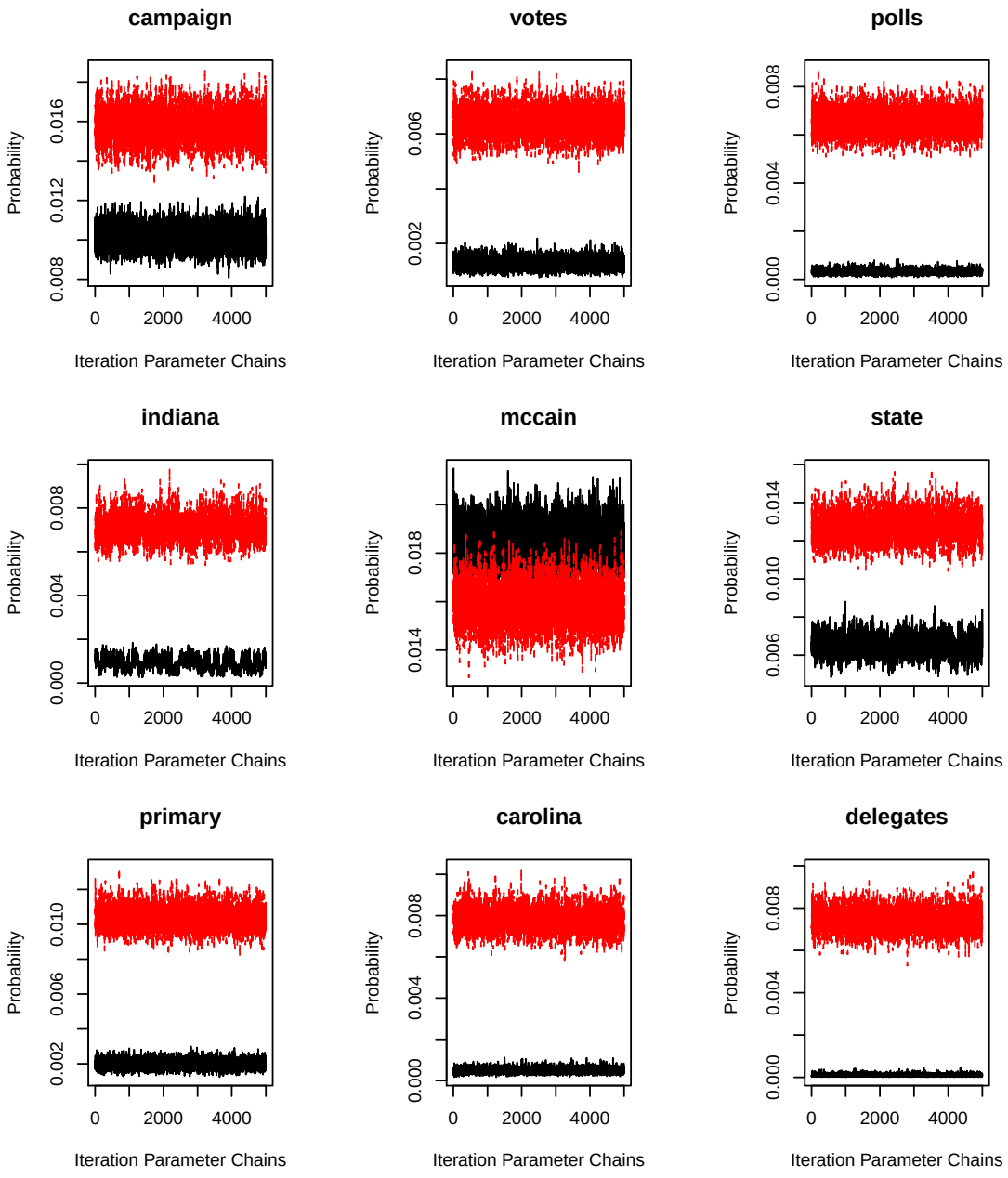


Figure B.12: MCMC results for the Political blog data with 10000 iterations, a 5000 burning period and a thinning of 10 iterations. MCMC output for the words, $\hat{\theta}_{jf}$, presented in Table 7.7.

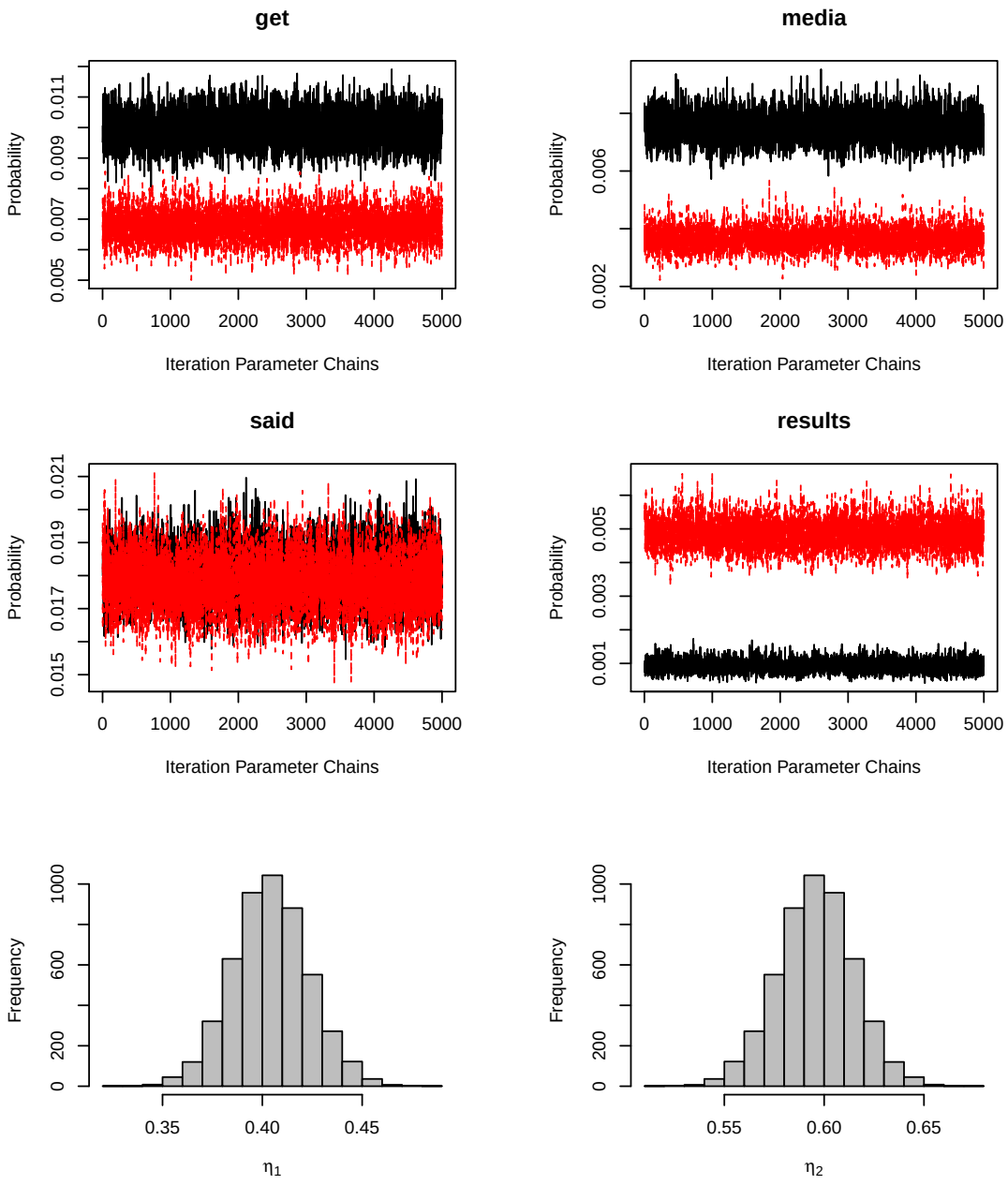


Figure B.13: MCMC results for the Political blog data with 10000 iterations, a 5000 burning period and a thinning of 10 iterations. MCMC output for the words, $\hat{\theta}_{jf}$, presented in Table 7.7 and output for weight parameters.

Appendix C

Probability density functions

- Multivariate Normal. $p(\mathbf{y} \mid \boldsymbol{\theta}_j) = \mathrm{N}(\mathbf{y}; \boldsymbol{\mu}_j, \Sigma_j)$ and $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j)$ where

$$p(\mathbf{y} \mid \boldsymbol{\mu}_j, \Sigma_j) = (2\pi)^{-p/2}|\Sigma_j|^{-1/2}\exp(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{y}-\boldsymbol{\mu}_j)),$$

  with $\mathbf{y} \in \mathbb{R}^p$, $\boldsymbol{\mu}_j \in \mathbb{R}^p$ and Σ_j a $p \times p$ symmetric positive definite matrix.

- Inverse-Wishart. $p(\Sigma_j \mid \nu, S) = \mathrm{IW}(\Sigma_j; \nu, S)$ where

$$p(\Sigma_j \mid \nu, S) = \frac{1}{C_{\Sigma_j}}|S|^{-\nu/2}|\Sigma_j|^{-(\nu+p+1)/2}\exp(-\frac{1}{2}\mathrm{tr}(S\Sigma_j^{-1})),$$

  with $C_{\Sigma_j} = (2^{\nu p/2}\pi^{p(p-1)/4}\prod_{f=1}^{p}\Gamma((\nu+1-f)/2))$, $\nu > p+1$ and S a $p \times p$ symmetric positive definite matrix.

- Gamma: $p(\kappa \mid a_\kappa, b_\kappa) = \mathrm{Gamma}(\kappa; a_\kappa, b_\kappa)$ where

$$\kappa \mid a_\kappa, b_\kappa) = \frac{b_\kappa^{a_\kappa}}{\Gamma(a_\kappa)}\kappa^{a_\kappa-1}\exp(-\kappa b_\kappa),$$

  with $a_\kappa > 0$ and $b_\kappa > 0$.

- Student-t: $p(\mathbf{y} \mid \boldsymbol{\theta}_j) = \mathrm{T}(\mathbf{y}; \boldsymbol{\mu}_j, \Sigma_j, \upsilon_j)$ where

$$p(\mathbf{y} \mid \boldsymbol{\mu}_j, \Sigma_j, \upsilon_j) = \frac{1}{C_t}|\Sigma_j|^{-1/2}\left(1+\frac{1}{\upsilon_j}(\mathbf{y}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{y}-\boldsymbol{\mu}_j))\right)^{-(\upsilon_j+p)/2},$$

with $C_t^{-1} = \dfrac{\Gamma((\upsilon_j + p)/2)}{\Gamma((\upsilon_j)/2)\upsilon_j^{p/2}\pi^{p/2}}$ and $\upsilon_j > 0$.

- Dirichlet: $p(\boldsymbol{\eta} \mid q) = \text{Dir}(\boldsymbol{\eta}; q)$ where

$$p(\boldsymbol{\eta} \mid q) = \frac{\Gamma(pq)}{(\Gamma(q))^p} \prod_{f=1}^{p} \eta_f^{q-1},$$

with $\eta_f > 0$ and $q > 0$.

- iskew-t: Consider the eigenvalue decomposition of the covariance matrix $\Sigma_j = A_j^{'} D_j A_j$, where $D_j = \text{diag}(d_{j1}, ..., d_{jp})$, $d_{jf} > 0$ is the j^{th} eigenvalue and $A_j \in \mathbb{R}^{p \times p}$ the non-singular eigenvector matrix. The j^{th} iskew-t component density is given by

$$p^{iskew-t}(\mathbf{y}_i|\boldsymbol{\theta}_j) = |\Sigma_j|^{-\frac{1}{2}} \prod_{f=1}^{p} \sum_{t=1}^{2} \frac{\Gamma(\frac{\upsilon_j+1}{2})}{\Gamma(\frac{\upsilon_j}{2})\sqrt{\upsilon_j \pi}} f_t(\mathbf{y}_i, \boldsymbol{\mu}_j, \upsilon_j, \alpha_{jf}^s, a_{jf}) \quad \text{(C.0.1)}$$

where

$$f_1(\mathbf{y}_i, \boldsymbol{\mu}_j, \upsilon_j, \alpha_{jf}^s, a_{jf}) = \left(1 + \frac{1}{\upsilon_j} \frac{(d_{jf}^{-1/2} a_{jf}^{'}(\mathbf{y}_i - \boldsymbol{\mu}_j))^2}{(1 - \alpha_{jf}^s)}\right)^{-1} \text{I}(d_{jf}^{-1/2} a_{jf}^{'}(\mathbf{y}_i - \boldsymbol{\mu}_j) \geq 0),$$

and

$$f_2(\mathbf{y}_i, \boldsymbol{\mu}_j, \upsilon_j, \alpha_{jf}^s, a_{jf}) = \left(1 + \frac{1}{\upsilon_j} \frac{(d_{jf}^{-1/2} a_{jf}^{'}(\mathbf{y}_i - \boldsymbol{\mu}_j))^2}{(1 + \alpha_{jf}^s)}\right)^{-1} \text{I}(d_{jf}^{-1/2} a_{jf}^{'}(\mathbf{y}_i - \boldsymbol{\mu}_j) < 0).$$

The specific component parameters in (C.0.1) are $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \Sigma_j, \alpha_{jf}^s, \upsilon_j)$ where α_{jf}^s and υ_j induce the asymmetry and tail thickness.

Bibliography

Affandi, R. H., Fox, E. B., and Taskar", B. (2013). Approximate inference in continuous determinantal processes. In *Advances in Neural Information Processing Systems*, pages 1430–1438.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6):3099–3132.

Andrews, G. E. (1998). *The theory of partitions*. Number 2 in 1. Cambridge university press.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2012). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*.

Berkhof, J., Mechelen, I. V., and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442.

Bianchini, I., Guglielmi, A., and Quintana, F. (2017). Determinantal point process mixtures via spectral density approach. *arXiv*, 1705.05181:1–42.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Biernacki, C. and Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, pages 451–457.

Brinkman, R. R., Gasparetto, M., Lee, J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R., and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 13(6):691–700.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673.

Chambaz, A. and Rousseau, J. (2008). Bounds for Bayesian order identification with application to mixtures. *The Annals of Statistics*, 36:928–962.

Chang, J. (2015). *lda: Collapsed Gibbs Sampling Methods for Topic Models*. R package version 1.4.2.

Chen, J. and Li, P. (2009). Hypothesis test for Normal mixture models: The EM approach. *The Annals of Statistics*, 37:2523–2542.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

Crawford, S. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89:259–267.

Dawid, A. (1999). The trouble with Bayes factors. Technical report, University College London.

Dempster, A., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, B*, 39-1:1–38.

Došlá, Š. (2009). Conditions for bimodality and multimodality of a mixture of two unimodal densities. *Kybernetika*, 45(2):279–292.

Drton, M. and Plummer, M. (2017). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23-1:1–22.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Frühwirth-Schnatter, S. (2006). *Finite Mixtures and Markov Switching Models*. Springer, New York.

Frühwirth-Schnatter, S. (2011). Dealing with label switching under model uncertainty. In Mengersen, K. L., Robert, C. P., and Titterington, M., editors, *Mixtures: estimation and applications*, volume 896, chapter 10, pages 213–239. John Wiley & Son.

Gassiat, E. and Handel, R. V. (2013). Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory*, 59(2):1115–1128.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Boca Raton: Chapman and Hall/CRC.

Ghosal, S. (2002). A review of consistency and convergence of posterior distribution. In *Division of theoretical statistics and mathematics*, pages 1–10, Indian Statistical Institute.

Ghosal, S. and der Vaart, A. V. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixture of normal densities. *Annals of Statistics*, 29:1233–1263.

Ghosal, S. and Van Der Vaart, A. (2007). Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35:697–723.

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In *Le Cam, L. M., Olshen, R. A. (Eds.), Proceedings of the Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer*, volume II, pages 789–806, Wadsworth, Monterey.

Gormley, I. C. and Murphy, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2(4):1452–1477.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for Normal mixture distributions. *The annals of Statistics*, 13:795–800.

Havre, Z. V., White, N., Rousseau, J., and Mengersen, K. (2015). Overfitting bayesian mixture models with and unknown number of components. *PLoS ONE*, 10 (7):1–27.

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44(6):2726–2755.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal Royal Statistical Society, B*, 72:143–170.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107:649–655.

Kan, R. (2006). From moments of sums to moments of product. *Journal of Multivariate Analysis*, 99:542–554.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20.

Lee, J. E. and Robert, C. P. (2016). Importance sampling schemes for evidence approximation in mixture models. *Bayesian Analysis*, 11:573–597.

Leroux, B. G. (1992). Consistence estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360.

Liu, X. and Shao, Y. Z. (2004). Asymptotics for likelihood ratio test in a two-component normal mixture model. *Journal Statistical Planing and Inference*, 123:61–81.

Lu, I.-L. and Richards, D. (1993). Random discriminants. *The Annals of Statistics*, 21:1982–2000.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295.

Marin, J. M. and Robert, C. P. (2008). Approximating the marginal likelihood in mixture models. *Bulleting of the Indian Chapter of ISBA*, 1:2–7.

Mengersen, K. L., Robert, C. P., and Titterington, D. M. (2011). *Mixtures: Estimation and Applications*. Wiley.

Mohsenipour, A. A. (2012). *On the distribution of quadratic expressions in various types of random vectors*. PhD thesis, The University of Western Ontario, Electronic Theses and Dissertation site.

Murphy, K., Gormley, I. C., and Viroli, C. (2017). Infinite mixtures of infinite factor analysers: nonparametric model-based clustering via latent gaussian models. *arXiv preprint arXiv:1701.07010*.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366.

Neal, R. M. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". *minmeo, University of Toronto*.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, 32(5):2044–2073.

Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897.

Ramamoorthi, R., Sriram, K., and Martin, R. (2015). On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789.

Ray, S. and Lindsay, B. (2005). The topography of multivariate normal mixtures. *The Annals of Statistics*, 33:2042–2065.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9:225–228.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixture models with an unknown number of components. *Journal of the Royal Statical Society*, B-59:731–792.

Rossell, D., Cook, J., Telesca, D., and Roebuck, P. (2018). *mombf: Moment and Inverse Moment Bayes Factors*. R package version 2.1.1.

Rossell, D. and Steel, M. F. J. (2017). Continuous mixtures with skewness and heavy tails. In Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P., editors, *Handbook of mixture analysis*, chapter 10. CRC press.

Rossell, D. and Telesca, D. (2017). Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265.

Rossell, D., Telesca, D., and Johnson, V. E. (2013). High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis*, pages 305–314, Springer.

Rousseau, J. (2007). Approximating interval hypotheses: p-values and Bayes factors. In Bernardo, J., Berger, J. O., Dawid, A. P., and Smith, A., editors, *Bayesian Statistics 8*, pages 417–452. Oxford University Press.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behavior of the posterior distribution in over-fitted models. *Journal of the Royal Statistical Society B*, 73:689–710.

Schork, N. J., Allison, D. B., and Thiel, B. (1996). Mixture distribution in human genetics. *Statistical Methods in Medical Research*, 5:155–178.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of statistics*, 6:461–464.

Shin, M., Bhattacharya, A., and Johnson, V. (2018). Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Teicher, H. (1963). Identifibility of finite mixtures. *The Annals of Mathematical Statistics*, 34:1265–1269.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*, volume 25 of *Cambridge monographs on applied and computational mathematics*. Cambridge University Press.

Watanabe, S. (2013). A widely applicable Bayesian information criteria. *Journal of Machine Learning Research*, 14:867–897.

West, M. and Turner, D. A. (1994). Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician*, 43:31–43.

Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2):415–428.

Xie, F. and Xu, Y. (2017). Bayesian repulsive Gaussian mixture model. *arXiv*, 1703.09061:1–85.

Xu, Y., Mueller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (dpp). *Biometrics*, 72(3):955–964.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematics and Statistics*, 39:209–214.